h e g

Enrichissement des dépôts institutionnels suisses : vers une couverture complète de la publication académique ouverte

Stratégie d'automatisation du moissonnage de plein-textes

Mémoire de recherche réalisé par :

Matthieu PUTALLAZ

Elodie SCHWOB

Sous la direction de :

Patrick RUCH, Professeur HES

Genève, le 17 janvier 2018

Master en Sciences de l'information

Haute École de Gestion de Genève (HEG-GE)

Hes·so GENÈVE
Haute Ecole Spécialisée de Suisse occidentale

Déclaration

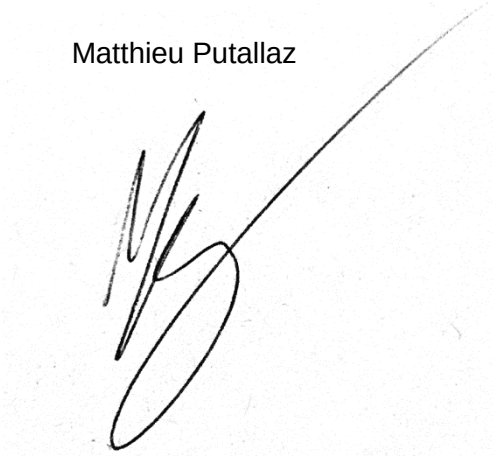
Ce mémoire de recherche est réalisé dans le cadre du Master en Sciences de l'information de la Haute école de gestion de Genève. L'étudiant accepte, le cas échéant, la clause de confidentialité. L'utilisation des conclusions et recommandations formulées dans ce travail, sans préjuger de leur valeur, n'engage ni la responsabilité de l'auteur/des auteurs, ni celle de l'encadrant.

« J'atteste/Nous attestons avoir réalisé le présent travail sans avoir utilisé des sources autres que celles citées dans la bibliographie. »

Fait à Genève, le 17 janvier 2018

Elodie Schwob

Matthieu Putallaz



Remerciements

Nous tenons à remercier premièrement notre responsable de projet Patrick Ruch, pour son suivi et ses remarques. Merci également à Julien Gobeill pour nous avoir si bien conseillé pour toute la partie modélisation et à Igor Milhit pour nous avoir expliqué les tenants et aboutissants du projet dans ses prémisses.

Merci à tous les assistants HES en Information documentaire qui ont lu notre poster et nous ont fait des remarques qui ont contribué à sa réussite et tout particulièrement à Nicolas Prongué pour ses critiques constructives. Nous tenons à remercier les personnes qui ont relu notre rapport final et qui l'ont corrigé.

Enfin, merci à nos collègues, amis et familles qui nous ont soutenus durant cette année.

Résumé

La publication d'articles scientifiques est le vecteur principal de la diffusion de l'information scientifique et de l'état d'avancement de la recherche. Les nouvelles possibilités de diffusion et de publication amenées par l'Open Access ont résolument modifié l'accès au contenu scientifique et positionnent les universités et les hautes écoles comme des acteurs incontournables dans la mise à disposition de la publication académique. En Suisse, ces institutions proposent des dépôts institutionnels qui ont la vocation de rendre accessible l'ensemble des publications de leurs chercheurs. Ces dépôts sont-ils complets ? S'ils ne le sont pas, est-il possible de créer une méthode automatique pour tendre à leur exhaustivité ?

Dans ce travail de recherche, nous proposons d'estimer le taux de couverture d'articles en plein-texte des dépôts institutionnels en les comparant aux résultats d'archives ouvertes de plus grande envergure. Ainsi, nous avons comparé la disponibilité des articles accessibles en plein-texte de sept dépôts institutionnels suisses (Edoc, Boris, ReroDoc, Serval, Archive Ouverte, Zora et Infoscience) à deux archives ouvertes (HAL et PMC). Les résultats de notre recherche indiquent que les dépôts institutionnels ne possèdent que 35,5% d'articles en plein-texte et qu'ils pourraient atteindre une couverture totale maximale de 74,8%. C'est donc l'opportunité, pour les institutions suisses, de plus de doubler leur couverture. Mais la diversité de licences des articles et la difficulté de les repérer est un réel défi pour les personnes qui gèrent ces dépôts.

Nous avons dans un deuxième temps modélisé une stratégie d'automatisation de la collecte de ces articles et en avons testé la faisabilité. La stratégie développée utilise Crossref et Open Access Button afin d'identifier les articles affiliés à une université et en trouver le plein-texte. Sur un échantillon de 141 notices, présentes en texte intégral à 43% sur deux dépôts test, notre stratégie permettrait de récolter 16% de plein-textes supplémentaires. Nous avons néanmoins rencontré un problème de taille dans l'identification des articles affiliés à des institutions spécifiques. En effet, dans le cas de l'Université de Lausanne, sur 200 notices testées, nous avons rencontré 36 appellations différentes alors que seules 4 appellations sont acceptées institutionnellement.

Mots-clefs : Dépôt institutionnel ; Libre accès ; Open Access ; Publication scientifique ; Articles scientifiques ; Archive ouverte ; Recherche scientifique ; Automatisation ; Plein-texte ; Texte intégral.

Table des matières

Liste des tableaux

Liste des figures

1. Introduction

L'information scientifique est principalement diffusée sous forme d'articles qui sont le résultat d'une démarche scientifique. Actuellement rendus disponibles de manière numérique, ces articles témoignent de l'avancement de la recherche et participent à la fois au rayonnement du chercheur qui les publie et de l'institution dans laquelle il travaille.

La modification des méthodes de diffusion des articles scientifiques a fortement impacté l'édition papier et permis l'émergence de l'Open Access. Les bibliothèques scientifiques renforcent alors leurs rôles et deviennent parties prenantes de l'archivage et de la mise à disposition des articles dans leur institution. Elles rejoignent le mouvement pour les archives ouvertes et le libre accès à l'information scientifique. Celui-ci s'oppose au modèle économique, proposé par les éditeurs et prend deux formes principales différentes :

- La publication d'un article dans une revue en libre accès, qui diffère des canaux commerciaux. L'ensemble de ces revues est répertorié dans le *Directories of Open Access Journal* (DOAJ) ;
- L'auto-archivage de l'article par son auteur dans une archive ouverte ou dans un dépôt institutionnel.

Une institution d'enseignement et de recherche peut contribuer à ce mouvement soit en créant un dépôt institutionnel et en incitant les chercheurs à y déposer leurs articles, soit en proposant des subsides aux groupes de chercheurs qui acceptent de publier dans une revue en libre accès (BIUM, 2017). Conscientes de l'enjeu essentiel de permettre une communication ouverte sur les résultats de recherches, les institutions suisses proposent presque toutes une plateforme d'archivage d'articles pour leur propre production.

La problématique qui en résulte est alors le risque d'un manque de visibilité pour un auteur dans le cas où tous ses articles ne seront pas disponibles sur les plateformes d'archivage institutionnelles. Le risque réside également dans le manque de visibilité de l'institution qui propose son dépôt institutionnel. C'est-à-dire qu'un manque trop prononcé d'articles ne serait pas le reflet exact des activités d'une institution.

Au cours de ce travail, nous allons donc analyser si ces dépôts institutionnels sont complets sur un plan national, dans quelle mesure, et s'ils ne le sont pas, pourquoi ?

Dans un deuxième temps, nous allons modéliser une stratégie d'automatisation qui permette d'identifier et de récolter automatiquement les articles issus de la recherche suisse. Le but étant de proposer une solution pour augmenter au maximum et légalement la quantité d'articles disponible dans les dépôts des institutions suisses.

1.1 Objectifs et questions de recherche

Dans le cadre du Master of Science en Sciences de l'information HES de la Haute école de gestion de Genève, nous nous sommes vus confier le projet de recherche suivant, dans son titre original : "*Acquisition des articles écrits par les chercheurs suisses à partir des archives institutionnelles internationales*".

Il s'agit d'évaluer la qualité des dépôts institutionnels dans les institutions de recherche et hautes études suisses ; universités, écoles polytechniques fédérales et les hautes écoles.

Toutes ces institutions de recherche sont à l'origine de nombreuses publications rédigées par des chercheurs. Elles proposent généralement ces articles sur une plateforme propre à l'institution. La communauté scientifique peut ainsi consulter les articles scientifiques rédigés dans une institution et permettre un accès ouvert et libre aux documents produits.

Toutefois, le postulat à l'origine de cette recherche est que ces différentes plateformes, ces dépôts institutionnels, ne sont pas complets. C'est à dire que tous les chercheurs écrivant des articles n'y déposent pas systématiquement leurs travaux. Les raisons sont multiples : embargo sur les délais de publication, respect des droits d'auteur, données sensibles, articles datant d'avant la création de l'archive institutionnelle ou simple oubli. Ces éléments participent à l'incomplétude de ces dépôts institutionnels, tant en ce qui concerne les textes intégraux que les notices bibliographiques.

La nature de ce mandat est une recherche exploratoire mixte qui nous amènera à quantifier les lacunes des archives institutionnelles en les comparants à d'autres bases documentaires (archives ouvertes, bases de données). Le but du projet est d'estimer et d'évaluer la quantité d'articles scientifiques en plein-texte disponibles dans les archives institutionnelles suisses, ainsi que d'évaluer la faisabilité d'une stratégie automatique d'acquisition du texte intégral.

1.1.1 Buts

- Quantifier la proportion de notices et d'articles en texte intégral présents dans les archives institutionnelles en les comparant à des sources de plus grande envergure ;
- Identifier des sources internationales qui hébergent des publications de chercheurs affiliés à des institutions suisses ;
- Développer et évaluer une méthode d'automatisation et d'identification du plein-texte.

Les buts ci-dessus nous ont permis d'émettre les trois questions de recherche suivantes :

1. Quelle est la proportion de publications produites par des chercheurs affiliés à des institutions suisses dans les dépôts institutionnels suisses ?

2. Comment identifier les articles issus de fonds de recherche suisse dans les sources internationales (archives ouvertes et plateformes commerciales) ?

3. Quelle part de la publication suisse absente des archives institutionnelles peut être obtenue automatiquement depuis les sources internationales et comment en systématiser l'acquisition ?

1.2 Définitions

Pour permettre la meilleure compréhension possible des termes que nous utilisons, nous proposons une définition du vocabulaire.

Dépôt institutionnel

Le dépôt institutionnel est un serveur sur lequel est déposée la production scientifique d'une institution. Les publications ainsi enregistrées sont accessibles au public et téléchargeables depuis cette plateforme. C'est l'institution qui se charge de faire l'archivage, qui finance et qui maintient son dépôt.

Archive ouverte

Une archive ouverte est une base de données sur laquelle est déposée une production scientifique. L'auteur ou l'ayant droit, si la licence le permet, dépose son article sur cette base qui est accessible librement par le biais d'internet. La différence entre le dépôt institutionnel et l'archive ouverte réside dans le fait que cette dernière n'est pas forcément rattachée à une institution : elle peut être internationale et ne traiter que d'un seul thème.

Le pré-print

Le pré-print est un article qui doit encore faire l'objet d'une relecture par des experts scientifiques du domaine et qui n'est par conséquent pas encore officiellement publié.

Le post-print

Par opposition au pré-print, le post-print est un article qui a fait l'objet d'une relecture et d'une validation par des experts scientifiques, mais qui n'est pas encore la version publiée de l'éditeur.

Les licences Sherpa/Romeo

Les licences de publication des revues scientifiques sont répertoriées sur le site Sherpa/Romeo[1]. Celui-ci propose une typologie des différentes licences et définit ce qu'il est possible de faire avec les articles.

Green : Il est possible de déposer sur une archive le pré-print et le post-print qui a déjà été officiellement publié et ce, dans sa version PDF officielle.

Blue : Il est possible de déposer sur une archive le post-print uniquement et la version déjà publiée et ce, dans sa version PDF officielle.

Yellow : Il n'est possible de déposer sur une archive que la version pré-print d'un article.

White : Le dépôt de l'article n'est pas possible.

[1] http://www.sherpa.ac.uk/romeo/search.php

Les API

API est l'acronyme de *Application Programming Interface*. C'est une interface proposée par les développeurs d'une application ou d'un site web, qui donne la possibilité à d'autres logiciels d'exploiter des données ou des fonctionnalités.

Le DOI

DOI est l'acronyme de *Digital Object Identifier*. Il s'agit d'un identifiant unique attribué à un article en ligne afin de faciliter son repérage.

Le potentiel d'augmentation de la couverture

Le potentiel d'augmentation de la couverture est un terme que nous utilisons pour cette recherche. Il s'agit de la proportion d'articles qui devraient se trouver dans les dépôts institutionnels en raison de leur licence et de leur provenance mais qui ne s'y trouvent pas.

2. Contexte général

2.1 Dépôts institutionnels

Les dépôts institutionnels sont des archives électroniques qui appartiennent à une institution et sont une vitrine de la production scientifique. Les chercheurs qui publient des articles dans le cadre d'une recherche sont tenus, en théorie, de déposer une version de leur article sur cette plateforme afin de, dans la limite de ce que permettent les licences, rendre ces articles disponibles librement sur cette plateforme. Ainsi l'institution qui finance la recherche peut disposer du résultat sans avoir besoin de racheter l'article auprès d'un éditeur.

C'est généralement la bibliothèque d'une institution qui prend la responsabilité d'une telle plateforme et qui la gère si elle en a les moyens financiers. Les avantages sont :

- Une meilleure visibilité de la production scientifique de l'institution ;
- Un atout lors de l'évaluation de celle-ci dans les divers classements mondiaux des universités ;
- Une économie pour les bibliothèques qui peuvent exploiter les articles de leur institution sans avoir besoin de les racheter auprès d'un éditeur.

Toutefois, il semble que ces types de dépôts ne soient pas complets et qu'une proportion considérable de ces articles ne s'y trouve pas, alors qu'elle devrait s'y trouver. Les raisons sont multiples :

- Le manque d'information de l'auteur, qui ne dépose pas systématiquement son article sur le dépôt institutionnel ;
- Les licences des journaux où sont publiés les articles, qui ne permettent pas le dépôt dans une archive ouverte ;
- Les articles avec une durée d'embargo qui ne sont pas débloqués et déposés dans le dépôt institutionnel après la période de verrouillage ;
- L'institution ou la bibliothèque n'ont pas les moyens de surveiller et de compléter ces dépôts quand cela est nécessaire.

2.2 Open Access

Les dépôts institutionnels étant un mode de distribution en libre accès, il est nécessaire de comprendre ce qu'est l'Open Access et ses enjeux.

L'Open Access, ou libre accès, est un mouvement qui vise le développement de l'accès et de la diffusion de l'information scientifique. L'Open Access s'est développé en réaction aux hausses, toujours plus fortes, des prix des licences d'abonnement aux revues scientifiques. Il offre une alternative à la publication scientifique classique en proposant un accès gratuit et libre aux articles. Les moyens de diffusion principaux de l'Open Access sont essentiellement les journaux qui publient selon un modèle libre, les réseaux sociaux scientifiques (Research Gate ou Academia) ou les dépôts institutionnels, qui font l'objet de ce travail.

Ainsi, pour pallier ce problème, plusieurs institutions et pays militent pour un accès facilité et plus libre à la production scientifique et visent à la dynamiser. Nous pouvons citer plusieurs actes qui font état de cette volonté, comme la « Déclaration sur l'accès aux données de la

recherche financée par des fonds publics » de l'Organisation de coopération et de développement économiques, l'Initiative de Budapest pour l'Accès Ouvert (BOAI), la déclaration de Berlin, et enfin la déclaration de Bethesda.

Cette dernière définit deux conditions afin qu'une publication soit en Open Access :

> *"1.Le/les auteur(s) ainsi que les titulaires du droit d'auteur accordent à tous les utilisateurs un droit d'accès gratuit, irrévocable, mondial et perpétuel et leur concèdent une licence leur permettant de copier, utiliser, distribuer, transmettre et visualiser publiquement l'œuvre et d'utiliser cette œuvre pour la réalisation et la distribution d'œuvres dérivées, sous quelque format électronique que ce soit et dans un but raisonnable, et ce à condition d'en indiquer correctement l'auteur ; ils accordent également aux utilisateurs le droit de faire un petit nombre de copies papier pour leur usage personnel.*
>
> *2.La version complète de l'œuvre, ainsi que tout document connexe, dont une copie de l'autorisation ci-dessus, réalisée dans un format électronique standard approprié, est déposée dès sa publication initiale dans au moins un réservoir en ligne subventionné par un établissement d'enseignement supérieur, une société savante, une agence gouvernementale ou tout autre organisme reconnu œuvrant pour le libre accès, la diffusion sans restriction, l'interopérabilité, et l'archivage à long terme (PubMed Central est un exemple de ce type de réservoir en sciences biomédicales)."*[2] *(Déclaration de Bethesda 2004)*

2.2.1 Modèles de publication en Open Access

Les deux modèles de publication principaux en OA sont :

La "**voie dorée**" ou gold OA : c'est lorsqu'un auteur publie directement son article dans une publication Open Access. L'auteur ou l'institution doit payer les frais de traitement. La liste de ces articles est disponible sur le répertoire des journaux disponibles en libre accès, le *Directory of Open Access Journal* (DOAJ).

La "**voie verte**" ou green OA : c'est lorsque l'auteur dépose son article (en pré ou post print, ou la version publiée) sur un système d'auto-archivage. C'est la voie qui est prise par les chercheurs qui déposent directement leur article sur le dépôt institutionnel de leur université.

2.3 Situation en Europe

La commission européenne, dans ses objectifs 2020[3], place le développement économique et l'innovation au premier plan. Elle pense qu'un accès privilégié à l'information scientifique est un levier important sur lequel agir pour influencer positivement le développement économique et scientifique. Pour cette raison, elle exprime une liste de recommandations afin de favoriser le libre accès aux articles scientifiques.

---

[2] http://openaccess.inist.fr/?Declaration-de-Bethesda-pour-l#n1

[3] http://ec.europa.eu/research/science-society/document_library/pdf_06/recommendation-access-and-preservation-scientific-information_en.pdf

2.4 Situation en Suisse

La situation en Suisse pour l'Open Access est en développement et il est soutenu largement par la conférence des recteurs des universités et des hautes écoles suisses : Swissuniversities. Il s'agit d'un comité de coordination nationale des différentes hautes écoles. En partenariat avec le Fonds national pour la recherche scientifique (FNS), publie une stratégie nationale suisse sur l'Open Access en 2017 et écrit :

> *"Une transition totale vers l'OA a le potentiel de contribuer positivement à la prospérité de la Suisse, prospérité qui repose principalement sur une formation de haute qualité, sur la recherche et l'innovation. Elle aura un fort impact non seulement sur les professions du savoir telles que la médecine, l'enseignement et le journalisme, mais aussi sur les petites et moyennes entreprises (PME)"*[4]

Cette stratégie repose sur des principes directeurs disponibles dans le document « Stratégie nationale suisse sur l'Open Access » (2017). C'est donc un but avéré d'opérer le plus efficacement possible une transition vers un modèle en Open Access.

2.5 Projet SONAR

Sur la base de cette stratégie nationale, il existe un pré-projet en réflexion à RERO (Réseau Romand des bibliothèques de suisse occidentale) : le projet Swiss Open Access Repository, ou SONAR. Il s'agit essentiellement d'évaluer la faisabilité de la création et de la mise en œuvre d'un dépôt institutionnel national qui rassemblerait tous les dépôts institutionnels. Les avantages d'un tel projet sont nombreux : on pourrait ainsi diminuer sensiblement les coûts de gestion des dépôts institutionnels pour les institutions, valoriser le travail scientifique de celles qui sont moins fortunées et qui n'ont pas les moyens d'exploiter un dépôt, centraliser et fédérer tous les articles sur une seule et même interface, dont la gestion reviendrait à RERO. Enfin, SONAR permettrait de simplifier le travail des bibliothèques scientifiques de Suisse.

3. Estimation de la couverture nationale

3.1 Remarques générales

Notre première question de recherche est la suivante : quelle est la proportion d'articles disponibles dans les dépôts institutionnels suisses accessibles en texte intégral ?

Afin d'estimer le taux de couverture des dépôts institutionnels suisses en matière d'articles en plein-texte, nous nous sommes basés sur des sources externes, comme par exemple des archives ouvertes, pour exploiter un nombre élevé d'articles en plein-texte affiliés à des institutions suisses.

En nous servant de ces sources comme d'un corpus documentaire de base, nous avons comparé les résultats de requêtes spécifiques sur ce corpus à ceux des dépôts institutionnels suisses. Cela nous a permis de compter le nombre d'articles que l'on ne retrouve pas dans les

[4] https://www.swissuniversities.ch/fileadmin/swissuniversities/Dokumente/Hochschulpolitik/Open_Access/Open_Access_strategy_final_f.pdf

dépôts, d'étudier les raisons de cette absence et d'estimer ainsi la couverture globale des dépôts suisses.

Notre démarche s'articule autour de :

- L'analyse et la sélection de sources d'articles (bases de données, archives ouvertes, etc.) ;
- La sélection de plusieurs dépôts institutionnels suisses afin d'atteindre un échantillon représentatif de la situation suisse ;
- La collecte manuelle des données dans les dépôts basés sur un échantillon de notices obtenu via les sources sélectionnées ;
- Le traitement et l'analyse des résultats permettant l'estimation de la couverture nationale.

3.2 Méthodologie

3.2.1 Choix des archives ouvertes

Dans un premier temps nous avons considéré un nombre important d'archives ouvertes et de dépôts internationaux, voire commerciaux, afin de sélectionner celui ou ceux qui pouvaient être les plus en adéquation avec nos exigences.

Nous avons considéré les sources suivantes :

- Arxiv, une archive de pré-prints dans le domaine des sciences dures, telles que la physique, les mathématiques, l'informatique etc.
- PubMed Central (PMC), une archive dédiée aux sciences de la médecine et aux sciences biomédicales. Près de 4,3 millions d'articles édités y sont disponibles.[5]
- SSRN, une archive électronique propriétaire de l'éditeur Elsevier qui propose plus de 700'000 articles dans 30 disciplines différentes.[6]
- SSOAR (Social Science Open Access Repository), une archive dédiée aux sciences sociales qui couvre plus de 30 journaux du domaine.[7]
- HAL (Hyper articles en ligne), est une archive ouverte française et pluridisciplinaire créée par le CNRS en 2001 qui héberge plus de 1'300'000 articles.[8]
- OAIster est une archive ouverte proposée par l'OCLC, qui met à disposition plus de 50 millions d'articles ou de notices bibliographiques.[9]

[5] https://www.ncbi.nlm.nih.gov/pmc/ [consulté en septembre 2017]

[6] https://www.ssrn.com/en/ [consulté le 17 mai 2017]

[7] http://www.ssoar.info/en/home.html [consulté le 17 mai 2017]

[8] https://hal.archives-ouvertes.fr/ [consulté le 19 décembre 2017]

[9] http://www.oclc.org/en/oaister.html [consulté le 11 janvier 2018]

- PloS, Public Library of Science, est une archive américaine, support de publication pour les articles sous licences libres. Elle contient essentiellement des articles de médecine et de biologie pour un total de 192'723 articles disponibles.[10]

Nous nous sommes également demandés s'il était possible d'exploiter des bases de données bibliographiques qui traitent uniquement les notices, comme par exemple Google Scholar, SciFinder ou Scopus. Mais comme notre but est de comparer les disponibilités des articles en texte intégral, nous n'avons pas poursuivi cette réflexion.

Pour faire un choix parmi ces archives, nous avons utilisé les critères de sélection suivants :

- La couverture thématique de l'archive ouverte ;
- Le nombre d'articles disponibles en plein texte ;
- La faisabilité de notre recherche.

Notre choix s'est arrêté sur deux archives ouvertes :

- PMC ;
- HAL.

Nous avons choisi ces deux archives car la première, PMC, est la plus complète en nombre d'articles (4,3 millions). Toutefois, parce que cette dernière est essentiellement dédiée aux sciences biomédicales, nous avons souhaité augmenter notre couverture thématique en sélectionnant HAL. Par ailleurs, nous avons exploré la possibilité d'utiliser Arxiv comme source d'articles. Malheureusement le repérage de la provenance et de l'affiliation des articles est difficilement réalisable. C'est-à-dire qu'il n'existe pas d'entrées pour une recherche sur Arxiv qui nous permette de filtrer les articles par affiliation. Comme il s'agit déjà d'une tâche minutieuse sur HAL et PMC, qui possèdent quant à elles un champ de recherche pour la provenance, nous n'avons pas sélectionné Arxiv.

3.2.2 Dépôts institutionnels et leurs échantillonnages

Les sources sélectionnées, nous pouvons commencer à faire un échantillonnage des dépôts institutionnels afin de démarrer la comparaison. En effet, nous avons pour cette phase procédé à deux échantillonnages stratifiés distincts. L'un au sein de l'ensemble des dépôts suisses et l'autre pour chacun de ces dépôts.

Pour commencer, nous avons sélectionné sept dépôts institutionnels sur les douze existants. Cette sélection s'est faite avec les critères de la représentativité nationale, la typologie des institutions, la distribution géographique et les domaines enseignés. Nous voulions représenter les hautes écoles, les universités et les écoles polytechniques. Nous avons procédé à l'évaluation de ces dépôts le 16 mai 2017. Les chiffres exprimés ici peuvent donc être différents aujourd'hui.

[10] https://www.plos.org/ [consulté le 11 janvier 2018]

Arodes par exemple est un dépôt institutionnel des HES. Il a été écarté en raison de la jeunesse de l'archive (créée en 2015), qui ne compte en mai 2017 que 1'367 notices ne concernant que le domaine Economie et service.[11]Le dépôt de l'Université, quant à lui, de Saint-Gall a été écarté en raison du manque de mise à jour de son site web[12].

Nous avons ainsi décidé de sélectionner les sept dépôts institutionnels suivants :

- **Edoc**, dépôt institutionnel de l'Université de Bâle, 43'314 notices ou articles disponibles ;[13]
- **Boris**, dépôt institutionnel de l'Université de Berne, 76'298 articles disponibles ;[14]
- **Réro Doc**, dépôt institutionnel des Hautes Ecoles de Suisse occidentale, de l'Université de Neuchâtel, de Fribourg et du Tessin. 19'997 notices ou articles y sont disponibles ;[15]
- **Serval**, dépôt institutionnel de l'Université de Lausanne et de l'Hôpital universitaire (CHUV), soit 124'466 notices ou articles disponibles ;[16]
- **Archive Ouverte**, dépôt institutionnel de l'Université de Genève et de l'Hôpital universitaire (HUG), soit 56'685 notices ou articles disponibles ;[17]
- **Zora**, dépôt institutionnel de l'Université de Zürich, soit 101'266 notices ou articles disponibles ;[18]
- **Infoscience**, dépôt institutionnel de l'Ecole Polytechnique Fédérale de Lausanne (EPFL), soit 129'498 notices ou articles disponibles.[19]

La somme de toutes les notices ou articles pour ces sept dépôts est de 551'524, que nous arrondissons à 552'000. C'est donc le chiffre que nous considérons comme étant le total (100%) des articles et notices disponibles sur les dépôts institutionnels.

A partir de cette sélection de dépôts, nous avons procédé à un échantillonnage représentatif des articles. Ainsi, nous avons choisi de faire une répartition proportionnelle par dépôt institutionnel sur une base de 0,1% du total des articles disponibles sur les dépôts. C'est-à-

[11] http://arodes.hes-so.ch/

[12] https://www.alexandria.unisg.ch/cgi/stats/report

[13] http://edoc.unibas.ch/cgi/stats/report

[14] https://boris.unibe.ch/cgi/stats/report

[15] https://doc.rero.ch/

[16] https://serval.unil.ch/

[17] https://archive-ouverte.unige.ch/

[18] http://www.zora.uzh.ch/cgi/stats/report

[19] https://infoscience.epfl.ch/ [consulté le 17 mai 2017]

dire que 0,05% des articles seront comparés à l'archive ouverte HAL et 0,05% à PMC, selon une clé de répartition identique à celle des articles dans les dépôts institutionnels.

Ceci représente donc 276 articles (0,05% de 552'000) répartis de la manière suivante :

- Edoc : 22 articles ;
- Boris : 38 articles ;
- Réro Doc : 10 articles ;
- Serval : 62 articles ;
- Archive Ouverte : 28 articles ;
- Zora : 51 articles ;
- Infoscience : 65 articles.

Nous effectuons la recherche sur la base de 276 articles de PMC et 276 articles de HAL, soit un total de 552 (0,1% de 552'000) articles différents.

Pour terminer l'échantillonnage, nous devons faire une sélection dans ce nombre d'articles :

- Les articles doivent être répartis également sur les années 2015 et 2016 ;
- Les articles doivent être triés du plus récent au plus ancien dans chacune de ces deux années.

3.2.3 Collecte des données

Concernant la collecte des données, nous avons utilisé le moteur de recherche de PMC et de HAL pour faire les requêtes qui nous permettent de sélectionner des articles de l'année 2015 et 2016, d'une des institutions. Puis, nous avons trié les résultats dans un ordre chronologique inversé.

Par exemple, voici un type de recherche que nous avons utilisé respectivement sur HAL et PMC, lorsque nous cherchions un article de l'Université de Berne pour l'année 2015 :

- HAL : Bern 2015 ;
- PMC : ("bern"[Affiliation]) AND "2015"[Date - Publication].

Nous avons multiplié ces requêtes pour chacune des institutions et des années souhaitées[20].

Ensuite, nous avons sélectionné les premiers articles disponibles en plein-texte selon notre clé de répartition. Une des difficultés rencontrées fut l'identification de l'affiliation des auteurs à l'institution concernée. Il a fallu vérifier tous les articles et s'assurer qu'au moins un auteur soit affilié à l'institution. Il est possible de le faire en consultant les métadonnées des auteurs sur la notice de l'article.

[20] La liste complète des requêtes est accessible sur demande.

De plus, nous avons vérifié la licence de chacun des articles sur Sherpa/Romeo. Enfin, nous avons fait une recherche sur les dépôts correspondants pour vérifier si les articles sélectionnés dans les archives ouvertes y étaient présents. Dans le cas où ils ne l'étaient pas ou pas entièrement, nous avons cherché à en connaitre la cause. C'est sur ce contrôle de licences systématique sur Sherpa/Romeo que nous nous sommes basés pour déterminer dans quelle mesure un article pouvait être diffusé sur un dépôt institutionnel.

Nous avons fait toutes nos recherches depuis des postes informatiques qui ne sont pas reliés à un réseau institutionnel pour éviter d'avoir accès à du contenu grâce aux licences mises à disposition par les institutions de recherche.

Pour chaque article, les données collectées sont les suivantes :

- Date de la requête ;
- Requête ;
- Noms des auteurs ;
- Nombre d'auteurs ;
- Titre ;
- Domaine thématique ;
- Langue de l'article ;
- Le statut de l'article (paru dans une revue, pré-print, document de travail) ;
- Nom de la revue ;
- Appellation de l'institution ;
- ID de l'auteur (si disponible) ;
- ID objet (DOI, HAL ID, PMC ID) ;
- Présence de la notice de l'article dans le dépôt institutionnel ;
- Présence de l'article en plein texte dans le dépôt institutionnel ;
- Licence (si mentionnée dans la source) ;
- Licence (si mentionnée dans le dépôt institutionnel) ;
- Type de licence sur Sherpa/Romeo ;
- Numéro de cas ;
- URL des dépôts et des archives ouvertes.

3.2.4 Classement des articles

Au cours de cette recherche, nous avons identifié plusieurs cas de figures différents puis classé les articles selon les différentes configurations rencontrées. L'objet de ces catégories est de déterminer les raisons pour lesquelles un article est présent ou absent sur un dépôt. Si un nouveau cas, inédit, apparaissait, nous rajoutions une nouvelle catégorie. Ainsi, nous avons identifié 18 cas différents. Nous avons mis en évidence les raisons pour lesquelles un article était présent ou absent du dépôt (type de licence, embargo, accès réservé aux membres de

la communauté scientifique), le type d'article (prépublication, publication revue par les pairs) et le type d'accès (notice seule, téléchargement en PDF ou lien vers la publication) tout en tenant compte des spécificités des dépôts institutionnels testés. La liste complète est disponible en **Annexe 1**.

3.3 Problèmes rencontrés

Le principal problème rencontré est celui de l'identification formelle de la provenance d'un article. Pour s'en assurer, nous avons vérifié l'affiliation de son auteur. Or, il est délicat de s'assurer que l'article en question est réellement censé se retrouver sur le dépôt institutionnel. En effet, dans le cas où plusieurs auteurs d'institutions différentes co-écrivent un article, dans quelle mesure un auteur, lié à une institution, est un critère suffisant pour affirmer que l'article qu'il a aidé à écrire doit se retrouver sur le dépôt institutionnel ? Au moins un auteur suffit-il ? Le problème est encore plus délicat dans le cas où un auteur possède plusieurs affiliations, car on ne sait pas par quelle institution il travaillait lors de la publication de l'article.

Un autre problème qui nous est apparu est celui de la couverture des archives HAL et PMC. En effet, nous avons considéré que toutes ces sources étaient plus complètes que les dépôts institutionnels. Mais la forte présence d'articles sous licences « green » nous invite à une certaine prudence avec ces résultats. Le nombre élevé de licences « green » est-il représentatif du type de licences utilisées lors d'un processus de publication en Suisse ?

Avec nos impératifs de recherche sur les sources d'articles HAL et PMC, nous n'avons parfois pas trouvé un nombre suffisant d'articles en texte intégral. C'est la raison pour laquelle sept articles manquent au total considéré (545 au lieu de 552 initialement prévu).

3.4 Résultats

Ci-dessous, voici les résultats que nous avons obtenus. Nous présentons en premier la couverture nationale et par dépôts testés en fonction de quatre catégories. Puis, nous détaillons le potentiel d'augmentation général et par dépôt. Ensuite, nous explorons les différents types de licences de publication avec les revues les plus représentatives. Enfin, nous présentons le détail des sujets des articles les plus traités dans nos recherches.

Nous avons classé les 18 cas rencontrés en quatre catégories :

A) **Présent** : l'article est présent sur le dépôt institutionnel (cas 1 ; 4 ; 5 ; 7 ; 13)
B) **Absent sans raison** : il s'agit ici du potentiel d'augmentation des dépôts institutionnels (cas 2 ; 3 ; 6 ; 15)
C) **Absent** : l'article est absent du dépôt pour des questions de licences (cas 8 ; 9 ; 11 ; 16 ; 17)
D) **Autre** : pas de type de licences trouvé (cas 12 ; 14 ; 18)

3.4.1 Couverture nationale

Pour ce qui est de la couverture des différents dépôts, voici ce que nous obtenons avec le corpus de 545 articles :

Tableau 1 : Couverture nationale et par dépôt

Dépôt	A) Disponible	B) Absent sans raison	C) Absent	D) Autre	TOTAL des absences
Boris	59.2%	26.3%	7.9%	6.6%	40.8%
Edoc	11.4%	72.7%	11.4%	4.5%	88.6%
Infoscience	26.9%	36.9%	33.1%	3.1%	73.1%
Serval	39.3%	41.8%	8.2%	10.7%	60.7%
Zora	43.0%	30.1%	20.4%	6.5%	57.0%
Rero Doc	25.0%	50.0%	12.5%	12.5%	75.0%
Archive Ouverte	25.0%	42.9%	21.4%	10.7%	75.0%
Total Suisse	**35.5%**	39.3%	18%	7.2%	**64.5%**

(L'ensemble des graphes par institution est disponible en **Annexe 2**.)

La couverture nationale s'élève à 35,5% d'articles disponibles dans les dépôts institutionnels suisses.

Nous constatons que le taux de disponibilité d'articles en plein-texte varie fortement selon le dépôt testé. La catégorisation des articles non disponibles permet également de confirmer qu'un taux non-négligeable de ceux-ci devrait se trouver sur les dépôts (absent sans raison).

On voit sur le tableau ci-dessus que tous les dépôts institutionnels ont un potentiel d'augmentation relativement important, allant de 26.3% pour Boris, le dépôt de l'Université de

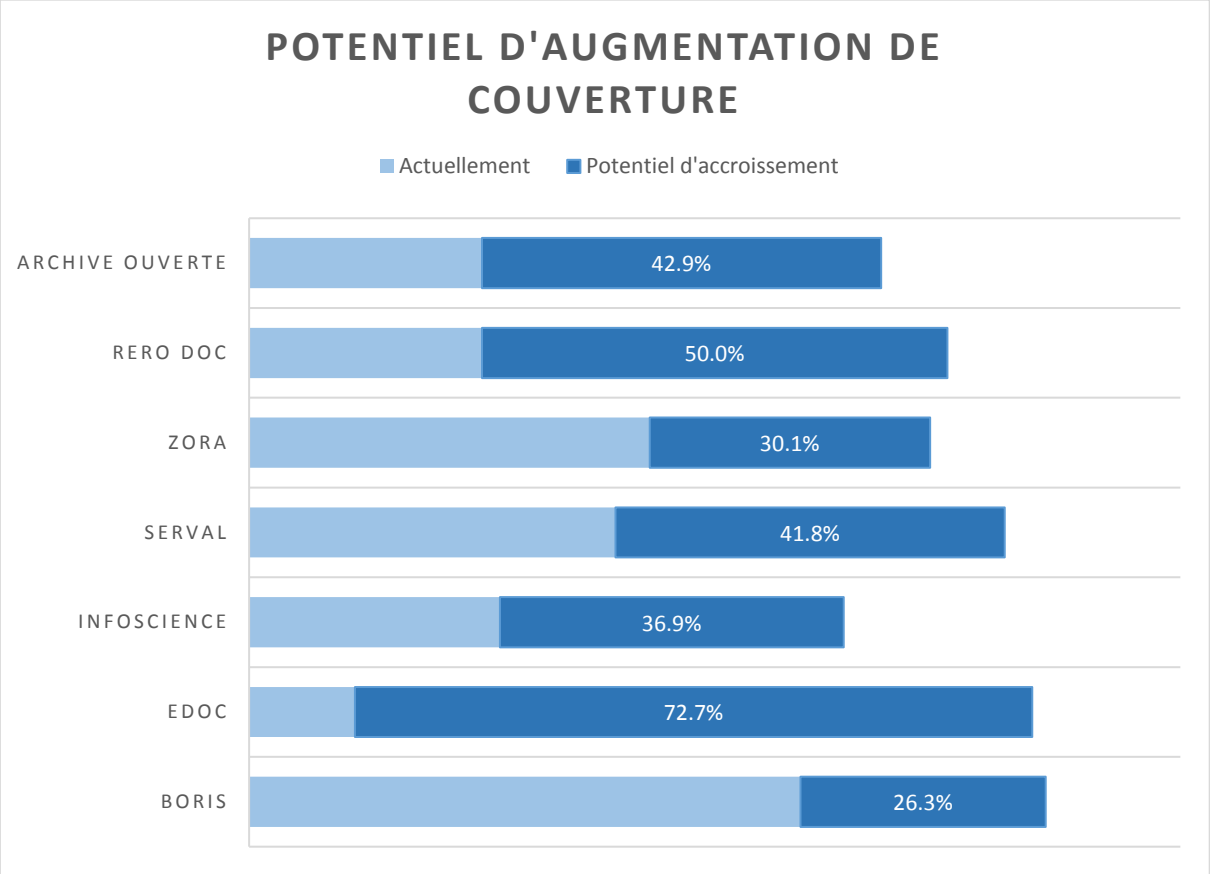
Berne, à 72.7% pour Edoc, le dépôt de l'Université de Bâle. Il est par ailleurs celui ayant la couverture la plus basse du panel.

3.4.2 Potentiel d'augmentation de la couverture

En nous basant sur la disponibilité des articles en plein-texte dont les licences permettent la diffusion, nous avons pu estimer le potentiel d'accroissement du taux de couverture par dépôt institutionnel.

En effet, une solution d'identification et de collecte automatique du plein-texte des articles issus de l'une de ces institutions permettrait d'augmenter le taux de couverture théorique jusqu'à 85.5%, dans le cas le plus optimiste.

Figure 1 : Potentiel d'augmentation de la couverture par dépôt



Pour ce qui est des différences pour les années 2015 et 2016, nous trouvons les résultats suivants :

Figure 2 : Couverture nationale pour les articles parus en 2015

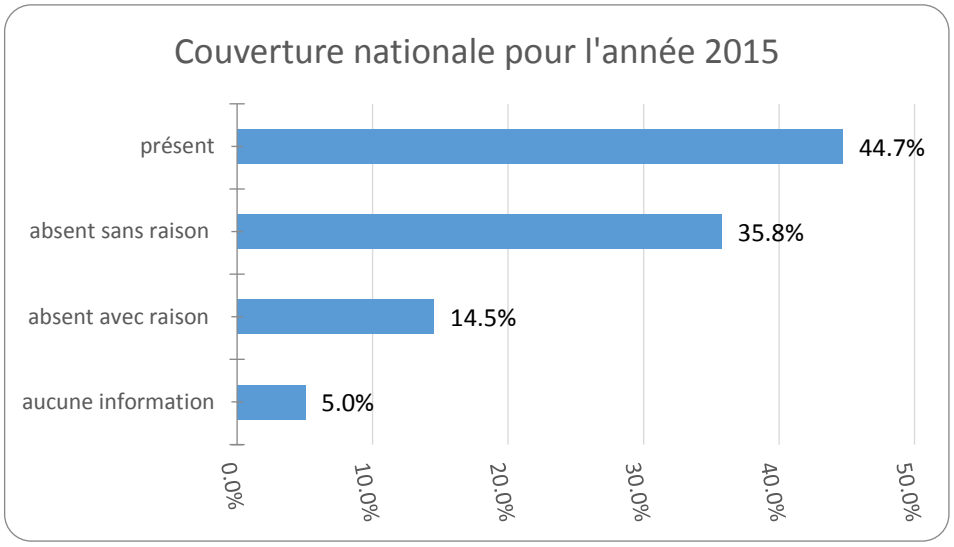
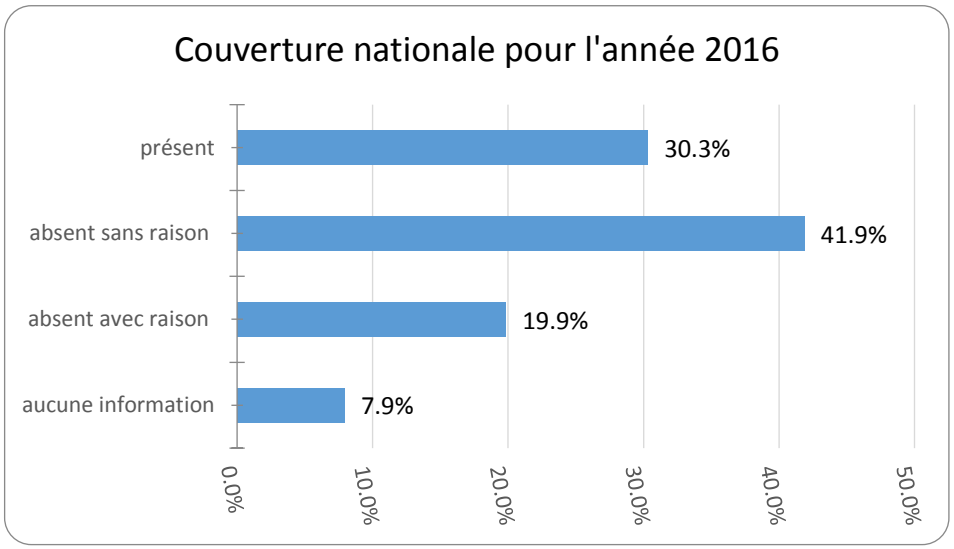


Figure 3 : Couverture nationale pour les articles parus en 2016



Comme nous pouvons le constater, la présence d'articles en plein-texte dans les dépôts testés est beaucoup plus importante en 2015 (44.7%) qu'en 2016 (30.3%). Effectivement, une partie conséquente des articles sous embargo passent après ce temps de verrouillage dans les archives ouvertes des institutions de recherche. La proportion de ce qui devrait être dans les dépôts institutionnels est donc plus importante en 2016. Vraisemblablement, ces articles n'ont pas encore eu le traitement nécessaire à leur intégration dans les dépôts institutionnels ou sont encore couverts par un embargo.

3.4.3 Point sur les revues

Nous avons utilisé la base de données Sherpa/Romeo afin d'identifier le statut des revues dans lesquelles les articles testés sont publiés. Sherpa/Romeo classe 2'430 revues (au 4 octobre 2017) dans les quatre catégories principales suivantes :

Figure 4 : Répartition du type de licences des journaux référencés sur Sherpa/Romeo

RoMEO colour	Archiving policy	Publishers	%
green	Can archive pre-print and post-print	1004	41
blue	Can archive post-print (ie final draft post-refereeing)	793	33
yellow	Can archive pre-print (ie pre-refereeing)	154	6
white	Archiving not formally supported	479	20

Summary: **80%** of publishers on this list formally **allow** some form of self-archiving.

(Sherpa/Romeo, 2017)[21]

Parmi ces quatre catégories, seule la catégorie « green » et sa sous-catégorie « green DOAJ » assurent le libre dépôt de l'article, son pré-print et son post-print sur un dépôt institutionnel. C'est pourquoi nous avons décidé que seuls les articles parus dans une revue couverte par ce type de licence seraient considérés comme devant logiquement se trouver sur le dépôt et donc compter dans le potentiel de croissance de disponibilité.

450 articles sur 545 (82.6%) ont été publiés dans des revues dont le nom est référencé dans Sherpa/Romeo. Toutefois, on observe que certains articles n'ont pas été publiés dans leur version définitive car nous avons également collecté des pré-print à partir de HAL uniquement, PMC n'offrant pas cette option.

Les revues francophones de sciences humaines sont les moins représentées dans la base de données.

[21] http://www.sherpa.ac.uk/romeo/statistics.php?la=en&fIDnum=|&mode=simple

Dans ce tableau, nous avons représenté le nombre d'articles pour chacune des licences Sherpa/Romeo :

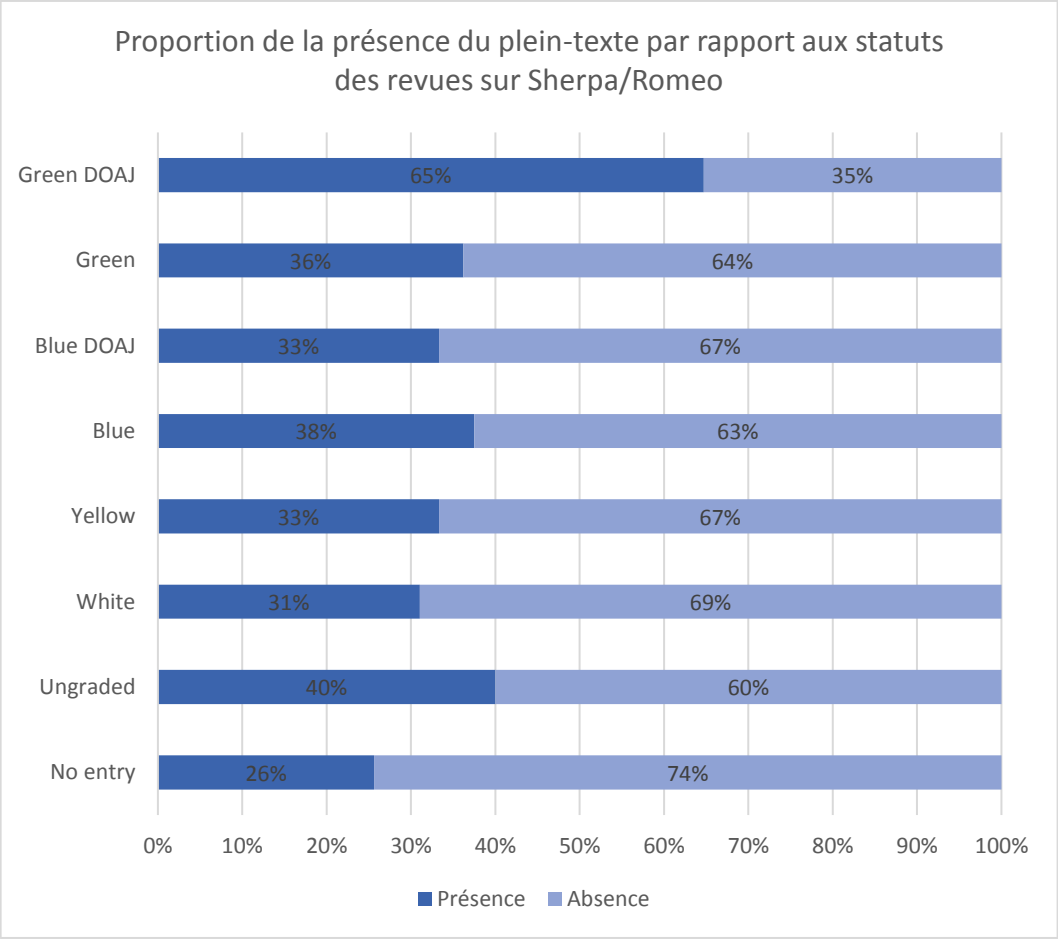
Tableau 2 : Répartition des articles par type de licences Sherpa/Romeo

Green DOAJ	Green	Blue	Yellow	White	Ungraded	Total
170	185	8	60	29	10	462

83 autres articles ne sont pas référencés dans Sherpa/Romeo.

Ci-dessous figure un tableau récapitulatif de la présence des articles plein-texte dans les dépôts par rapport au statut de la revue sur Sherpa/Romeo :

Figure 5 : Proportion de la présence du plein-texte par rapport aux statuts des revues sur Sherpa/Romeo



Nous remarquons que, normalement, nous devrions trouver un taux de 100% pour ce qui est de des revues qui bénéficient du statut « green » et « green DOAJ ». Ce qui représente également un potentiel d'accroissement pour les dépôts étudiés si nous observons seulement les licences.

Tableau 3 : Occurrence des revues

N°	Journal	Editeur	Sherpa/ Romeo	Plein-texte	Pas de plein-texte
27	PLoS one	Public Library of Science	Green DOAJ	15	12
13	Scientific Reports	Nature publishing Group	Green DOAJ	6	7
9	Swiss medical weekly	EMH Swiss Medical Publishers	Green DOAJ	8	1
7	Nature communications	Nature Publishing Group	Green DOAJ	5	2
6	Toxicology In Vitro	Elsevier	Green	1	5
4	Astrophysical Journal	American Astronomical Society	Green DOAJ	2	2
4	Monthly Notices of the Royal Astronomical Society	Oxford University Press	Green	0	4
4	eLife	eLife Sciences Publications	Green DOAJ	3	1
4	Frontiers in aging neuroscience	Frontiers Media	Green DOAJ	4	0
4	Chemistry		Ungraded	2	2

On remarque que les revues qui ont publié la majorité des articles collectés sont des journaux offrant une option Open Access. Il est intéressant de constater que même pour les revues les plus fréquemment trouvées, une part importante des articles en plein-texte ne se trouve pas dans les dépôts.

3.4.4 Domaines les plus représentés

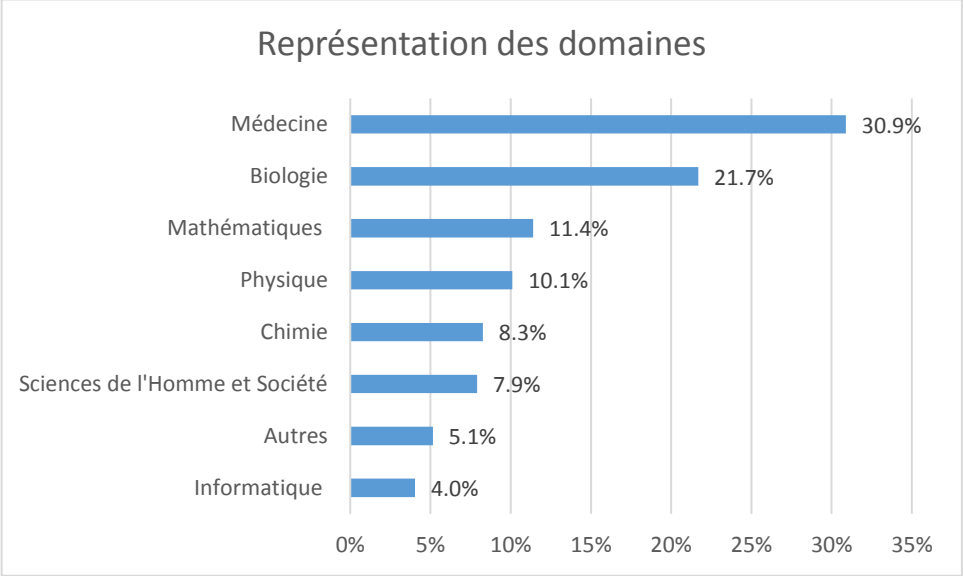
Dans notre échantillon, nous avons travaillé sur les articles des domaines suivants :

- Biologie ;
- Botanique ;
- Chimie ;
- Economie ;
- Environnement ;
- Informatique ;
- Mathématiques ;
- Médecine ;
- Pharmacologie ;
- Physiologie ;
- Physique ;
- Psychologie ;
- Sciences cognitives ;

- Sciences de l'environnement ;
- Sciences de l'homme et société ;
- Sciences de l'information ;
- Sciences de l'ingénieur ;
- Sciences sociales ;
- Sociologie ;
- Toxicologie.

Voici la répartition des domaines sur l'ensemble des articles :

Figure 6 : Répartition des domaines d'étude des articles



On remarque une surreprésentation de la médecine (30.9%) car il s'agit du domaine principal couvert par PMC qui est l'une des deux sources, avec HAL, des articles recherchés dans les dépôts institutionnels suisses correspondants.

On peut constater cette différence sur la répartition des domaines par source :

Figure 7 : Répartition des domaines sur PMC

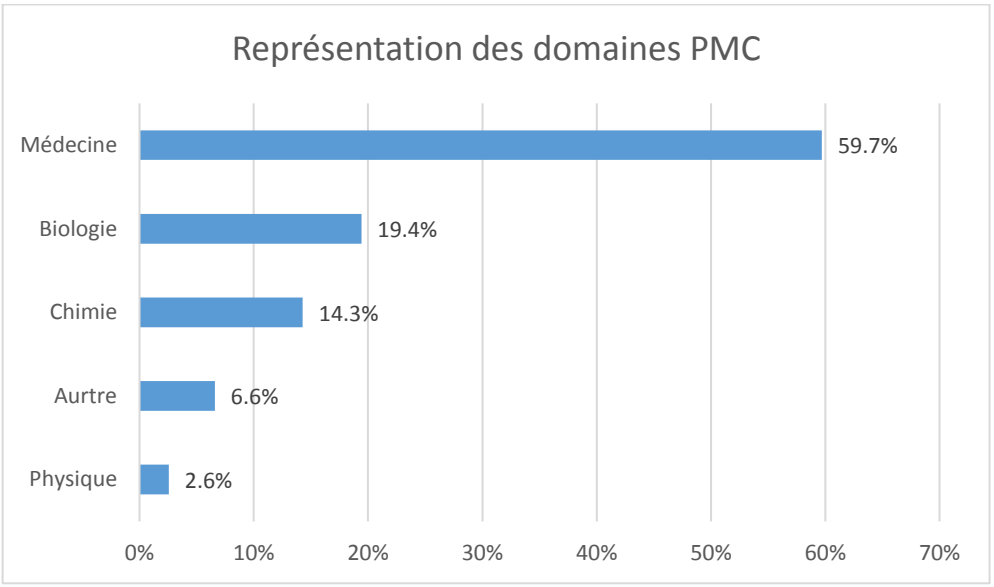
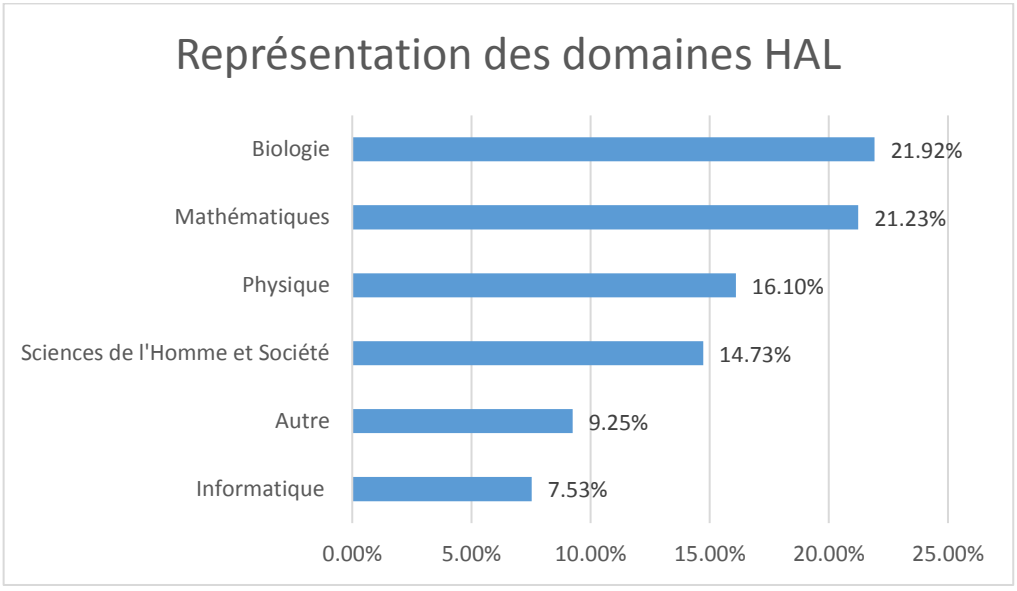


Figure 8 : Répartition des domaines sur HAL



On note une répartition plus homogène sur HAL que sur PMC, HAL étant une source pluridisciplinaire. Toutefois, la présence importante des sciences de la vie et de la médecine, disciplines pionnières en OA (AUBRY, 2005), se retrouve dans les deux archives.

3.5 Synthèse des résultats

En l'état, nous remarquons que la couverture globale des dépôts institutionnels suisses est de 35,5% et qu'il serait possible de la doubler pour l'amener à 74,8%, soit un potentiel d'augmentation de 39,3%. S'il peut s'agir d'un chiffre élevé, il est à nuancer car toutes les institutions n'ont ni les mêmes moyens, ni les mêmes priorités en matière de dépôts institutionnels. Cependant, ce chiffre n'est pas idéal pour la mise en valeur des articles des chercheurs et des institutions qui ne bénéficient pas d'une visibilité représentative.

Nous sommes conscients de certaines limites de notre recherche et les résultats obtenus sont à considérer avec prudence. Ci-dessous figurent les principales limites que nous avons identifiées :

- Pour des raisons de métadonnées disponibles dans nos sources, nous avons considéré le champ “affiliation” comme suffisant pour déterminer la provenance du financement de l’article ;
- Il nous est impossible, dans le cas où un chercheur a plusieurs affiliations, de savoir avec exactitude quelle est l’affiliation déterminante permettant de connaître la provenance du financement ;
- Le choix de HAL et PMC, qui hébergent principalement des articles en sciences dures, entraîne une sous-représentation statistique des sciences humaines et sociales ;
- La majorité des articles que nous avons étudiés sont publiés dans des revues Open Access car les sources que nous avons choisies favorisent ce mode de publication ;
- Les articles déposés dans HAL ou PMC y ont été déposés par des chercheurs déjà conscients de l’importance du partage de leur travail dans des archives ouvertes ;
- La grande majorité des articles testés sont en licence libre « green » et « green DOAJ », de ce fait, l’ensemble des résultats n’est pas représentatif de la publication scientifique dans sa globalité.

3.6 Identification des articles suisses

La seconde question de recherche concernait l’identification des articles issus de fonds de recherche suisses dans les sources internationales (archives ouvertes et plateformes commerciales).

Ainsi, les 540 articles testés sur HAL et PMC ont tous fait l’objet d’une vérification manuelle. En effet, l’adresse de l’affiliation de l’auteur figure dans la mention de responsabilité de chacun des articles. Grâce à cette information, nous avons pu identifier la provenance des articles. Certaines fois l’information apparaissait clairement sur la notice : par exemple, « Université de Genève ». D’autres fois, nous avons dû déduire l’affiliation depuis une adresse : par exemple, « 30 quai Ernest-Ansermet ch-1211 Genève 4 » qui est l’adresse de la section des sciences de l’Université de Genève. Dans ces cas de figure, il a fallu s’assurer que les adresses correspondaient à l’institution que nous cherchions. Les problématiques des appellations sont abordées dans les chapitres 4.4 et 4.5 de ce rapport.

4. Stratégie d'automatisation

4.1 Remarques générales

Après avoir calculé le potentiel d'augmentation de la couverture des dépôts institutionnels suisses, nous avons développé une stratégie permettant d'automatiser la récolte des articles en plein-texte manquants.

Il s'agit de mettre en place un processus à l'aide de plusieurs outils et plateformes déjà existants qui permette, à l'aide d'un script, d'acquérir automatiquement tous les nouveaux articles en plein-texte et ce dès leur mise à disposition sur des sources extérieures à l'institution dans le cas où ils n'auraient pas été déposés sur le dépôt institutionnel.

Nous cherchons ici à répondre à la question de recherche n°3 :

Quelle part de la publication suisse absente des archives institutionnelles peut être obtenue automatiquement depuis les sources internationales et comment en systématiser l'acquisition ?

Notre démarche s'articule autour de :

- La modélisation de ce processus ;
- La comparaison des différents outils existants ;
- La sélection d'une combinaison d'outils ou de plateformes qui permette la mise en application de la stratégie ;
- Un test manuel de notre stratégie sur un échantillon aléatoire afin de calculer le potentiel concret de récupération de plein-texte de la stratégie.

4.2 Modélisation de la stratégie

4.2.1 Concepts envisagés

Dans le cadre des recherches préliminaires de notre projet de recherche, nous envisagions une récolte de plein-texte via les différentes sources d'envergure existantes, soit, par exemple, directement sur HAL ou PMC. Or, lors de nos différentes recherches, nous avons réalisé que de nombreux outils permettant de trouver le plein-texte existaient déjà.

Durant notre travail, nous avons été confrontés à plusieurs reprises à la frustration de chercheurs face à la faible couverture d'articles scientifiques en plein-texte que leurs institutions leur offraient (soit au travers d'un dépôt institutionnel, soit au travers des abonnements aux revues payantes fournis par la bibliothèque de leur institution). Nous nous sommes ainsi mis à réfléchir à un usage institutionnel de Scihub. Il s'agit d'un site web qui héberge de nombreux articles scientifiques en plein-texte sous licences libres et propriétaires gratuitement, mais souvent obtenus illégalement.

L'étude sur le taux de couverture de Scihub de Himmelstein (2017), constate ainsi que sa couverture est presque complète selon les domaines : elle atteint par exemple 92.8% en chimie. Toutefois, devant la controverse autour de l'usage de cet outil et les différentes

réactions des éditeurs propriétaires à l'encontre de membres d'institutions qui en faisaient l'apologie (PEET, 2016), nous avons décidé de nous focaliser sur des solutions moins disruptives. Nous avons néanmoins choisi de conserver Scihub dans notre panel de test afin de comparer notre solution à cet outil.

4.2.2 Analyse des outils

Dans le but de déterminer l'outil d'identification de plein-textes déjà existant le plus exhaustif en matière de sources, nous avons élaboré une grille d'évaluation en **Annexe 3**.

Les outils évalués ont été identifiés pendant l'état des lieux du projet. Pour les évaluer, nous avons retenu les critères suivants :

- Nom ;
- Date de création ;
- Gouvernance ;
- Type de solution ;
- Fonctionnement ;
- Domaines couverts ;
- Nombre d'articles totaux ;
- Taux de retour de plein-texte (si connu) ;
- Sources des articles ;
- Présence d'une API ;
- Uniquement Open Access ou mixte ;
- Recherche par DOI.

Nous avons ainsi comparé onze outils présélectionnés car ils permettent de retrouver le plein-texte d'un article.

Nous avons finalement choisi d'utiliser Open Access Button, principalement en raison du nombre de sources différentes interrogées, de la présence d'une API (en français, Interface de programmation applicative), de la recherche dans tous les domaines et qu'elle soit possible depuis le DOI de l'article.

4.2.3 Limites techniques et légales

Nous avons été contraints d'adapter notre stratégie en fonction des restrictions légales et techniques que nous avons rencontrées.

En matière de contrôle des licences des publications dans lesquelles sont parus les articles testés, nous avions utilisé jusque-là Sherpa/Romeo. Malheureusement, les métadonnées du site web, accessibles via son API, sont peu fournies et donc difficilement exploitables. Nous sommes ainsi partis de l'idée qu'Open Access Button permettrait un tri de licences satisfaisant. Or, nous nous sommes aperçus que ce n'était pas toujours le cas. OAB permet, certes, dans la majorité des cas, de collecter des articles légalement disponibles mais ce n'est pas sa raison d'être : il donne accès à des articles disponibles en plein-texte, ce qui n'est pas synonyme de

légalement disponible. Nous avons néanmoins décidé qu'il ferait office de filtre, malgré la marge d'erreur possible.

La seconde limite que nous avons rencontrée fut à propos de l'utilisation de Scihub. Au-delà des considérations éthiques qui entourent l'usage institutionnel de cette base de données, nous nous sommes concentrés sur la faisabilité technique de son utilisation dans un script.

Nous avons rencontré deux problèmes importants :

Le premier sont les captcha (*Completely Automated Public Turing test to tell Computers and Humans Apart*) qui apparaissent après trois ou quatre requêtes dans la même heure et empêchent ainsi les machines d'avoir accès au PDF. Cela signifie qu'il faudrait déterminer exactement sur combien de requêtes et en quel lapse de temps ils apparaissent pour les éviter.

Le second est relatif à l'instabilité des noms de domaines utilisés par Scihub. En effet, le site, illégal dans plusieurs pays, subit régulièrement des fermetures de domaines. Au cours de notre test, qui dura trois semaines, nous avons utilisé un seul accès à Scihub (sci-hub.io), mais ce dernier fut inaccessible à de nombreuses reprises plusieurs heures durant. Ce domaine a d'ailleurs été fermé depuis notre test.

Nous avons donc renoncé à utiliser Scihub dans notre stratégie mais l'avons conservé à titre comparatif dans le test afin de se faire une idée de la différence de couverture entre un outil légal et Scihub.

4.2.4 Concept retenu pour le test

Après l'évaluation des différents outils et la prise en compte des limites légales et techniques de notre projet de stratégie, nous avons modélisé notre stratégie d'automatisation afin de la tester sur un échantillon de notices.

Nous avons choisi comme point de départ les DOI des articles plutôt que leur titre. Le DOI est un identifiant unique attribué à un objet numérique qui permet sa gestion pérenne.

4.2.4.1 Crossref

Nous avons choisi d'utiliser Crossref comme source des nouveaux articles publiés affiliés aux institutions suisses.

Crossref est une agence d'enregistrement de DOI. Elle domine le marché de l'attribution de DOI en Occident en étant responsable de 60% des DOI attribués dans le monde, mais de 99.9% des DOI présents sur les liens des pages Wikipédia en anglais (KIKKAWA, 2016).

Les métadonnées des articles (en format JSON) sont directement accessibles et interrogeables via l'API de Crossref et permettent d'identifier l'institution d'affiliation et le fonds de subvention de chaque article, ainsi que son éditeur et la revue dans laquelle il est publié :

Exemple de métadonnées :

Fonds de subventions :

"funder":[{"name":"Universit\u00e9 de Gen\u00e8ve (CH)","award":["Universit\u00e9 de Gen\u00e8ve"]}]

Editeur :

"publisher": *"Springer Nature"*

Nom de la revue :

"container-title":["Human Brain Mapping"

"short-container-title":["Int J Semiot Law"]

4.2.4.2 Les dépôts institutionnels

Une fois les DOI obtenus sur Crossref, nous avons vérifié la disponibilité en plein-texte des articles sur les dépôts institutionnels correspondants. En l'occurrence, nous avons sélectionné l'Université de Lausanne et celle de Genève en raison de leur taille et de leurs facultés relativement similaires ainsi qu'en raison du potentiel d'accroissement de leurs dépôts institutionnels se situant pour les deux autour de 40%.

4.2.4.3 Open Access Button

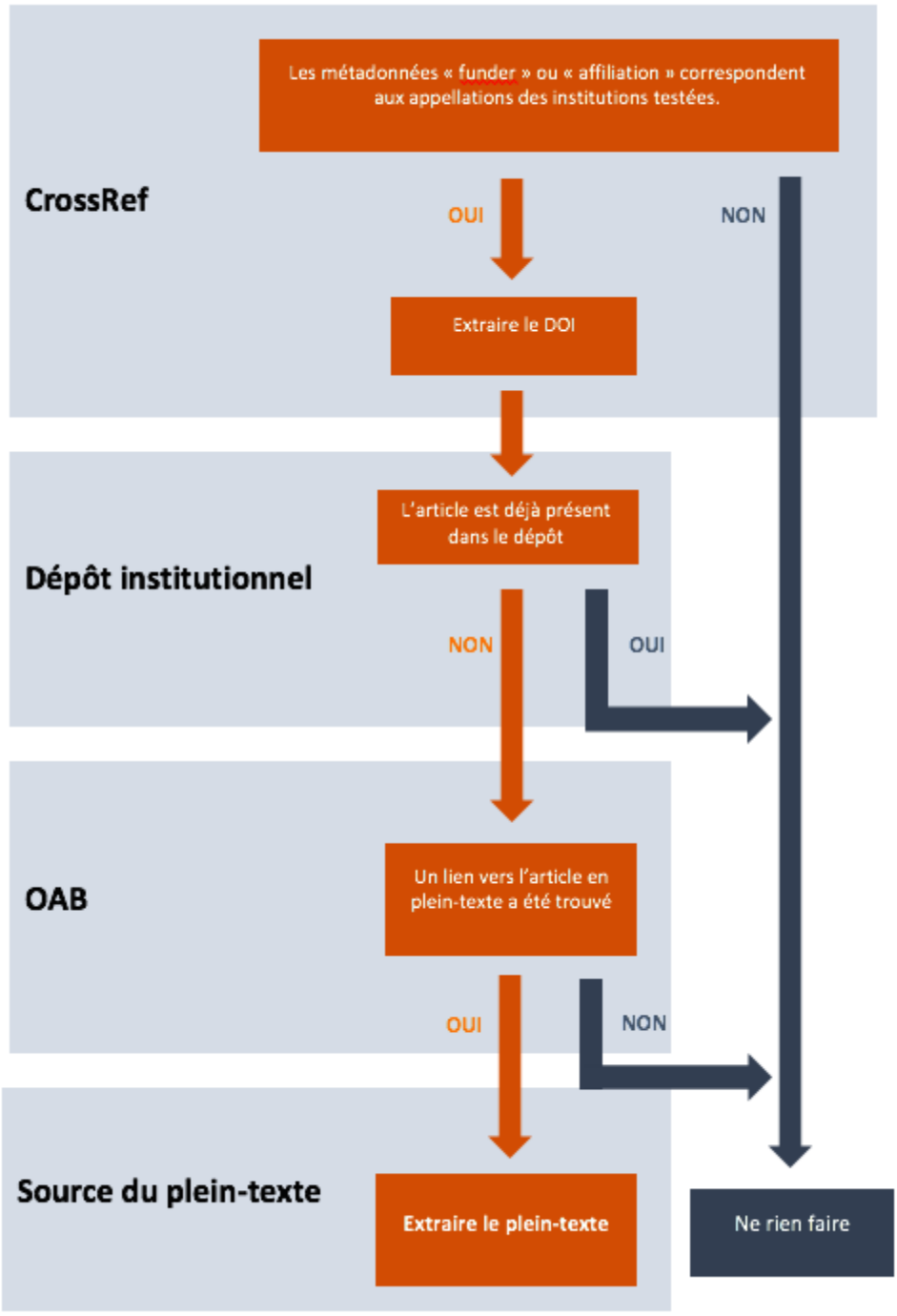
Lorsque l'article n'est pas disponible en plein-texte, nous interrogeons alors OAB, toujours à partir du DOI de l'article.

Open Access Button est une extension de navigateur ainsi qu'un site web permettant de rechercher le plein-texte d'un article scientifique de manière légale dans des milliers de sources différentes.

Il permet de rechercher un article depuis, entre autres, son DOI, son titre ou une URL. S'il trouve l'article, OAB fournit un lien qui renvoie sur la source permettant l'accès au document PDF ou directement au PDF. Lorsqu'il ne trouve pas l'article, il peut le demander directement à son auteur.

4.2.4.4 Modélisation graphique

Figure 9 : Modélisation de la stratégie d'automatisation de la collecte



4.3 Test de la stratégie

4.3.1 Méthodologie

L'objectif du test de notre stratégie de moissonnage est d'estimer l'augmentation effective du taux de couverture qu'elle pourrait entraîner pour les dépôts institutionnels et si ce taux est significatif. Ce test a eu lieu entre le 2 et le 21 novembre 2017.

Crossref

Afin d'obtenir 70 DOI d'articles affiliés à l'Université de Lausanne et 70 affiliés à l'Université de Genève, nous avons procédé à un échantillonnage de type probabiliste aléatoire.

Le nombre de DOI retenus a été défini pour des raisons de faisabilité, soit du temps à disposition et de la capacité du moteur de recherche de Crossref à retrouver les articles affiliés à l'Université en question (les champs "funders" et "affiliation" n'étant pas obligatoires et rarement renseignés).

Pour les deux universités, nous avons retenus trois termes de recherche différents et une couverture de quatre années différentes :

Lausanne :

"université lausanne", 2016

"lausanne university", 2015

CHUV, 2017 et 2014

Genève :

"université genève" 2016

"geneva university" 2015

HUG 2017 et 2014

Les résultats obtenus ont été classés par le critère de pertinence (relevance) et nous avons sélectionné tous les articles jusqu'à obtenir 70 DOI de chaque université.

A partir des DOI, nous avons créé un tableau Excel dans lequel collecter les données suivantes :

- Titre de l'article ;
- DOI ;
- Présence de l'article sur le dépôt (oui, notice seule, non) ;
- Disponibilité de l'article par OAB (lien, PDF, non) ;
- Source sur laquelle OAB nous renvoie ;
- Code de la page permettant d'identifier l'élément PDF ;
- Présence d'une API ;
- Présence de l'article sur Scihub.

Les dépôts institutionnels : Serval et Archive Ouverte

Nous avons sélectionné les deux plus grandes universités romandes, soit l'Université de Lausanne et son dépôt Serval, ainsi que l'Université de Genève et son dépôt Archive Ouverte, en raison de la diversité des domaines étudiés et de la faculté de médecine qui y est rattachée dans les deux cas.

Une fois les DOI obtenus via Crossref, nous avons interrogé les dépôts institutionnels correspondant afin de vérifier la présence du plein-texte de l'article. Nous avons effectué la requête à partir du DOI mais lorsque nous n'obtenions pas de résultats nous vérifions une seconde fois le dépôt avec une requête par titre de l'article.

Open Access Button

Nous avons ensuite vérifié la disponibilité du plein-texte de l'article par OAB. Dans un premier temps, nous avons interrogé OAB uniquement par le DOI des articles.

Cela nous a retourné cinq types de résultats différents :

1. L'article n'a pas été trouvé mais peut être demandé à l'auteur.
2. L'article n'est pas disponible, une demande a été faite à l'auteur mais cette dernière a été refusée.
3. L'article n'est pas encore disponible, une demande a été faite à l'auteur et est pour l'instant, sans réponse.
4. L'article est disponible via la plateforme de l'éditeur, le dépôt institutionnel ou l'archive ouverte.
5. L'article est disponible directement en PDF.

Nous n'avons pas différencié les cas 1 à 3 dans la collecte des données : les trois cas sont regroupés dans la catégorie non disponibles.

Dans le cas 2 où l'article était disponible via une source externe, nous avons identifié cette source ainsi que l'élément du code de la page permettant d'identifier le PDF rattaché à celle-ci. Il serait ainsi possible d'écrire un script pouvant extraire le PDF de toutes les sources rencontrées.

Nous avons également vérifié la présence d'une API sur la page source car cela faciliterait la collecte automatique du PDF.

Suite à la session poster ayant eu lieu le 14 décembre, nous avons décidé dans un second temps de réitérer les recherches sur OAB des articles non disponibles. En effet, nous nous sommes aperçus qu'en recherchant uniquement par DOI, nous avions exclu les pré et post-prints des résultats potentiels, ces derniers n'ayant pas de DOI attribués.

Nous avons donc répété le test en nous servant cette fois des titres des articles.

Le tableau Excel des résultats du test a donc été modifié en ajoutant le critère suivant :

- Présence de l'article sur OAB : recherche par titre.

Scihub

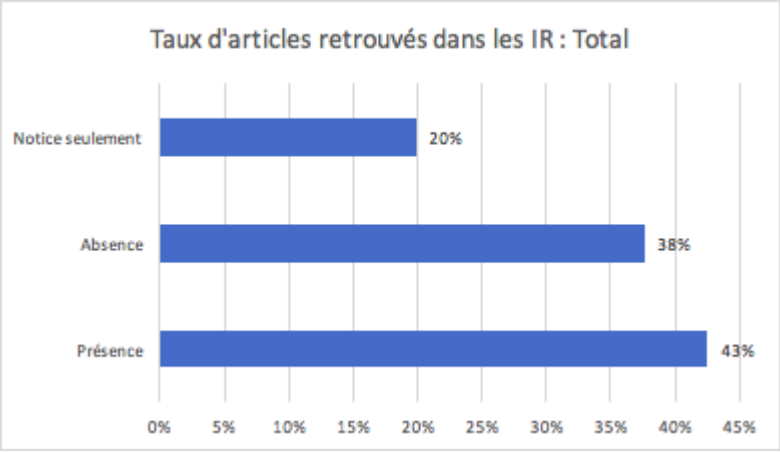
La dernière étape de notre test consiste en la vérification de la disponibilité de l'article sur Scihub.

Nous avons utilisé un des nombreux noms de domaine de Scihub (sci-hub.io, désormais fermé) et effectué des recherches par DOI, qui renvoient directement l'article en PDF sans intermédiaire.

4.3.2 Analyse des résultats de la stratégie de moissonnage

Vérification de la disponibilité du plein-texte sur les dépôts institutionnels des DOI identifiés sur Crossref

Figure 10 : Répartition des DOI recherchés dans les dépôts



Seuls les articles non disponibles nous intéressent car ce sont eux qui vont nous permettre d'augmenter la couverture des dépôts.

Ce premier tri nous permet de concentrer nos recherches de la seconde étape sur les 20% de notices seules et les 38% d'absence totale des dépôts. La deuxième étape du test de la stratégie se concentre sur les 82 DOI (58%) dont le plein-texte est absent des dépôts.

Proportion des notices seules

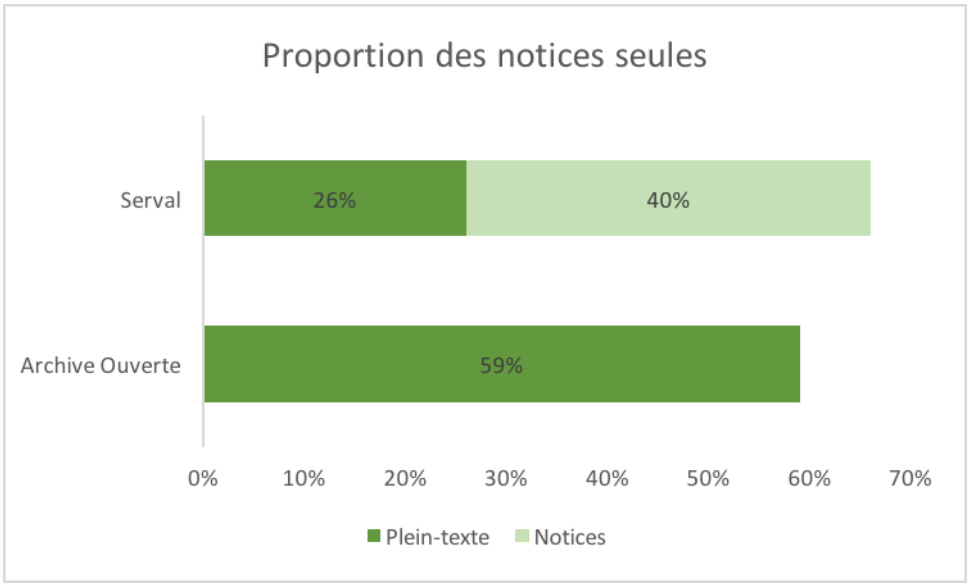
Il est intéressant de noter la grande disparité entre Serval et Archive Ouverte en matière de disponibilité du plein-texte. En effet, sur Serval seuls 26% des articles des DOI testés sont disponibles en plein-texte, contre 59% sur Archive Ouverte. Ces chiffres s'expliquent par une gestion différente des articles sous licences propriétaires.

A l'Université de Genève les articles qui ne sont pas en accès libre sont disponibles via un login au personnel et aux étudiants de l'Université. La présence de l'article ne signifie donc pas forcément son accessibilité à tous.

L'Université de Lausanne a choisi une autre approche en ne proposant pas une gestion d'accès à plusieurs niveaux aux articles déposés. Les articles n'ayant pas de licence libre n'y sont pas déposés et seule leur notice y est disponible.

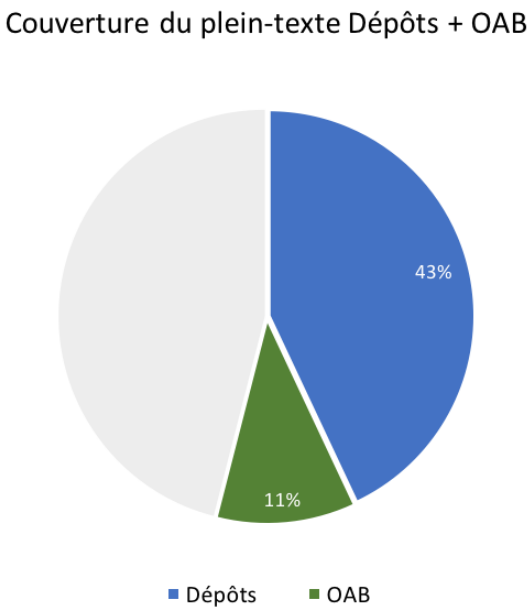
De ce fait, si on rajoute les notices au pourcentage « disponibles » de Serval, on arrive à une valeur équivalente de 66%. Néanmoins, comme nous ignorons si les articles dont seule la notice est disponible sont déjà possédés par Serval, nous avons été obligés de les compter comme non-disponibles.

Figure 11 : Proportion de notices seules dans Serval



Calcul du nombre d'articles non disponibles sur Serval et Archive Ouverte dont le plein-texte est récupérable grâce à Open Access Button

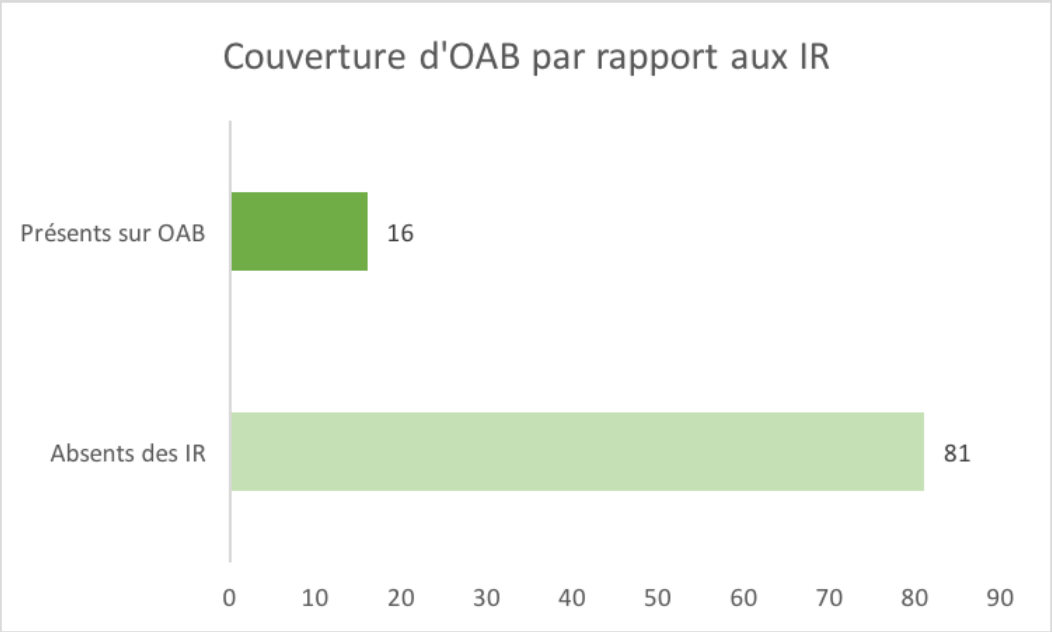
Figure 12 : Couverture du plein-texte sur les dépôts et Open Access Button



OAB permet d'accroître le taux de couverture des deux dépôts de 43% à 54%, soit une augmentation de 11% des articles en plein-texte disponibles sur le dépôt.

Cela représente 20% des articles qui manquent sur les dépôts institutionnels

Figure 13 : Couverture d'Open Access Button par rapport aux dépôts



Cela signifie qu'encore 46% de la totalité des DOI ne sont pas récupérables avec Open Access Button. Nous avons développé plusieurs hypothèses permettant d'expliquer ce chiffre :

- Nous estimons à environ 20% les articles parus sous licence propriétaire et qui ne permettent pas leur diffusion libre. Le taux de 46% peut ainsi être ramené à un taux de 26% de manque dans la couverture.
- Certaines métadonnées des sources "crawlées" (par exemple InfoScience ou HAL) sont mal comprises par OAB qui interprète le renvoi vers le fulltext comme un chemin libre vers le plein-texte alors qu'il renvoie régulièrement à la version payante de l'article.

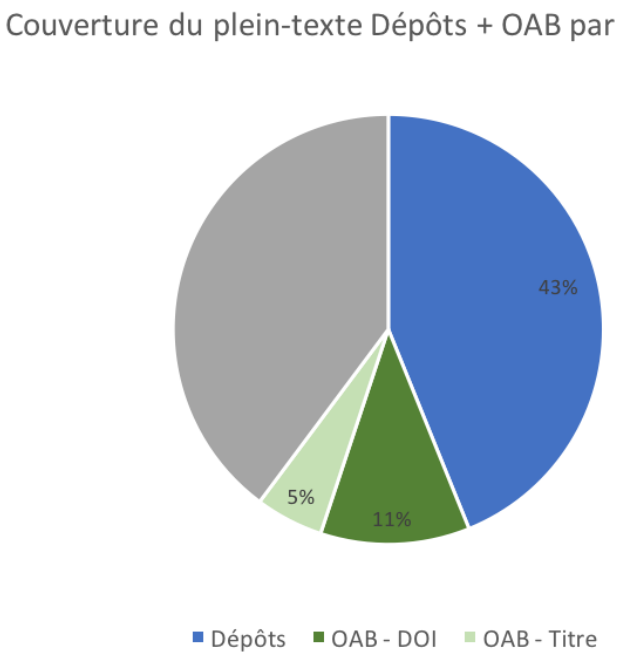
Figure 14 : Exemple d'erreur de liens sur Open Access Button

La recherche des DOI seule exclut les pré et post-prints des articles, ces derniers n'ayant pas de DOI attribués.

La recherche par titre sur OAB

Afin de contrer le biais évoqué ci-dessus nous avons décidé de répéter la recherche des articles sur OAB mais à partir d'une requête par titre d'article et non par DOI.

Figure 15 : Couverture du plein-texte sur les dépôts et avec Open Access Button en recherche par titre

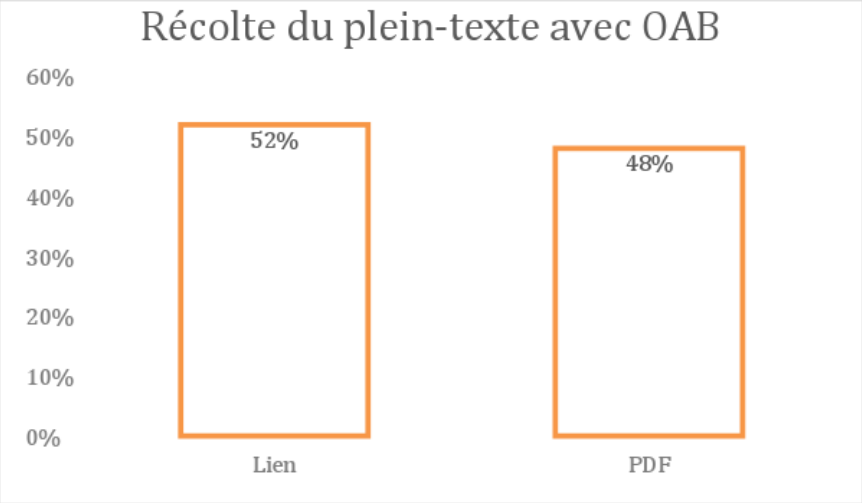


La recherche par titre sur OAB a permis de récolter 8 articles supplémentaires sur les 141 DOI de notre échantillon de base. Il est à noter qu'un seul de ces 8 articles était un post-print et ne disposait ainsi pas d'un DOI. Les autres articles étaient en version éditeur mais n'avaient pas été identifiés par leur DOI pour une raison indéterminée.

4.3.2.1 Récolte du plein-texte par Open Access Button

Open Access Button permet de récupérer un lien pointant vers un site qui héberge le PDF ou directement sur le PDF. Nous avons donc différencié les résultats en fonction de la solution proposée par l'outil.

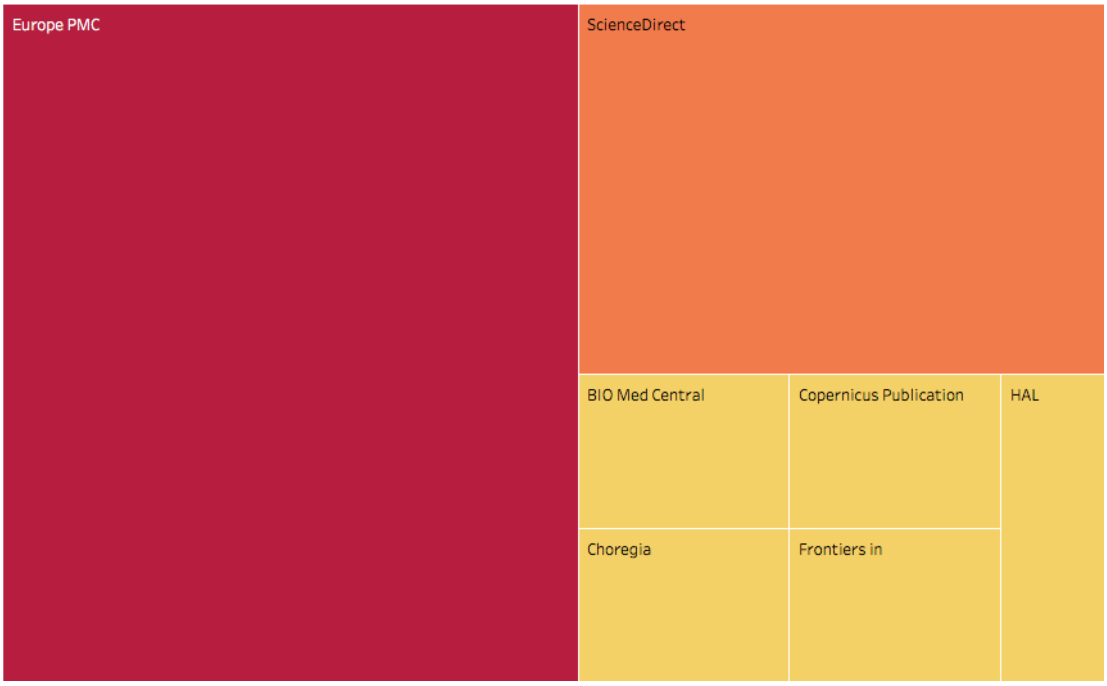
Figure 16 : Répartition des types de réponses positives sur Open Access Button



Les liens qui pointent directement sur le PDF sont les plus facilement traitables. Il suffit de télécharger automatiquement le PDF dès que l'on arrive sur la page du document. Cela se complique lorsqu'OAB pointe sur un site qui possède le PDF. Dans ce cas, nous sommes obligés d'aller dans le code source de la page afin de comprendre comment les métadonnées sont structurées pour pouvoir mettre dans notre script tous les cas rencontrés pour moissonner le PDF hébergé sur la page.

Pour ce faire, nous avons listé les différentes plateformes (sites d'éditeurs, archives ouvertes et dépôts institutionnels) sur lesquelles OAB nous renvoie et identifié les métadonnées pointant sur le PDF du plein-texte de l'article.

Figure 17 : Répartition des différentes plateformes sur lesquelles renvoie Open Access Button



Parmi les 23 liens donnés par OAB, deux sources représentent le 78% de la provenance des plein-textes. Il s'agit de Europe PMC, base de données biomédicale gratuite qui donne accès, en décembre 2017, à 4,5 millions[22] d'articles en plein-texte et de ScienceDirect[23], plateforme d'Elsevier, qui permet parfois d'avoir accès au plein-texte d'articles gratuitement une fois la période d'embargo terminée.

[22] http://europepmc.org/

[23] http://www.sciencedirect.com/

Les métadonnées identifiant l'élément PDF de ces deux sources sont :

- Europe PMC

* PDF *

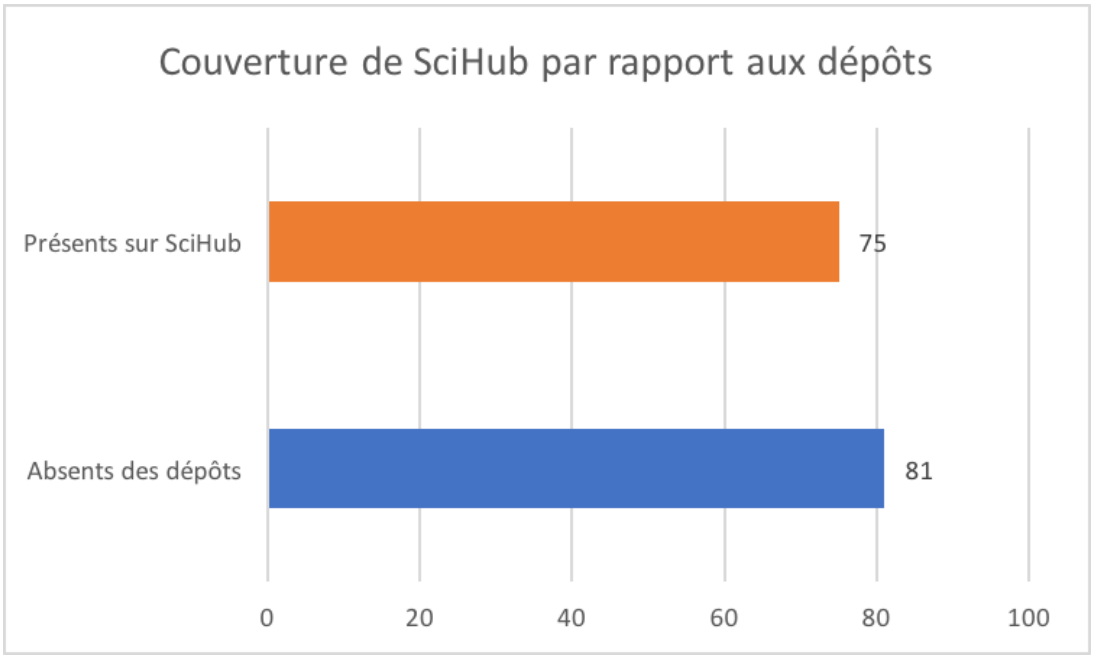
- ScienceDirect

<meta data-react-helmet="true" name="citation_pdf_url" content="/science/article/pii/S0012160616303979/pdfft?md5=12bb5891aa2d07da8e132db557838d39&pid=1-s2.0-S0012160616303979-main.pdf"/>

Notre script devrait ainsi réussir à identifier les différentes méthodes que les sites sources ont pour mettre à disposition le PDF.

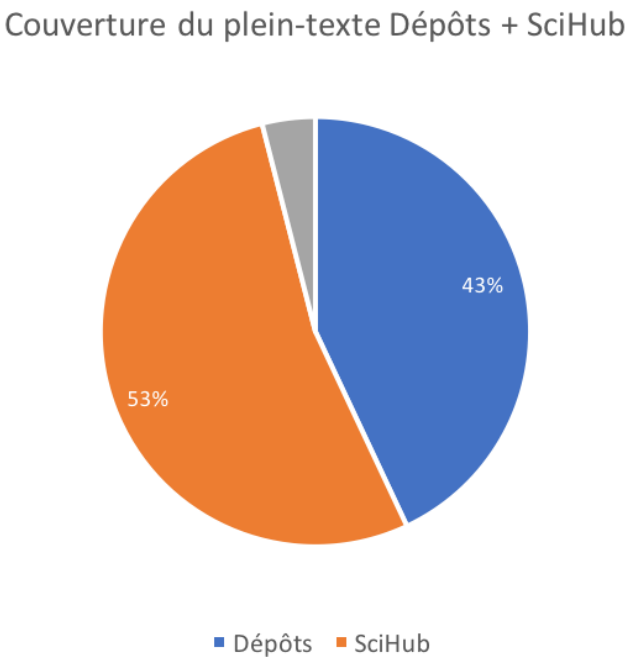
Comparaison des résultats obtenus avec Open Access Button avec ceux de Scihub

Figure 18 : Couverture de Scihub par rapport aux dépôts



93% des articles manquants des dépôts sont disponibles en plein-texte avec Scihub. Cela représente une augmentation effective du taux de couverture des dépôts de 53%.

Figure 19 : Couverture totale des dépôts et de Scihub



En utilisant Scihub nous arrivons à une couverture quasi complète des articles en plein-texte. Seuls 4% des plein-textes des articles restent inaccessibles.

4.3.3 Synthèse des résultats de la stratégie

Des 39.3% de potentiel d'augmentation envisagé dans la partie 3 : Estimation de la couverture nationale de ce projet, notre proposition de stratégie d'automatisation permettrait un taux d'augmentation de 16% pour atteindre une couverture globale de 51.5%.

En matière d'outils, nous nous sommes aperçu qu'Open Access Button était moins efficace qu'attendu. Nous avons rencontré de nombreuses erreurs de liens qui pointaient vers la version payante de l'article. Par rapport à Scihub, qui permet d'obtenir 93% des articles manquants, OAB n'en ramène que 20%.

L'autre problème que nous avons rencontré dans la mise en place de notre stratégie fut l'identification dans Crossref des articles écrits par des chercheurs affiliés aux universités test. Nous allons développer cette difficulté dans le chapitre suivant.

4.4 Le cas des appellations

4.4.1 Problématique

Nous avons été confrontés au problème des différentes appellations des institutions tout au long de ce projet. D'abord, en cherchant sur HAL et PMC des articles affiliés aux sept universités testées, nous nous sommes rendu compte de l'importante diversité des appellations sans que cela ne nous entrave dans cette première étape. Ce n'est qu'au cours du test de la stratégie d'automatisation de la récolte que nous avons été freinés par ce problème.

En effet, en recherchant les articles rattachés aux Université de Lausanne et de Genève nous nous sommes aperçus que peu d'articles répondaient à une appellation standardisée et nous avons été contraints d'élargir nos recherches à la ville de l'université, ce qui nous a forcé à effectuer un important travail de tri, les articles affiliés à l'EPFL, par exemple, se trouvant également dans les résultats de "Lausanne".

En nous basant sur une étude réalisée sur les articles publiés par les chercheurs de l'Université Lyon Part Dieu en 2005 (BADOR et LAFOUGE 2005), nous avons effectué notre propre test sur les différentes appellations rencontrées dans deux universités romandes : l'UNIL et l'UNIGE.

4.4.2 Collecte des appellations

Nous avons choisi l'Université de Lausanne et celle de Genève car il s'agit des deux universités les plus importantes de Suisse romande auxquelles sont rattachés un hôpital universitaire :

- L'UNIL, 14'475 étudiants[24], 7 facultés, CHUV, HEC
- L'UNIGE, 16'530 étudiants[25], 9 facultés, HUG

Nous avons choisi d'effectuer le test à partir de HAL, archive ouverte sur laquelle nous avions déjà eu l'occasion de travailler dans la première partie du projet de recherche et dont les champs "affiliations" des auteurs sont généralement bien renseignés et peu normalisés. Nous avons également sélectionné ScienceDirect, plateforme de l'éditeur Elsevier qui est multidisciplinaire (ce qui est important pour la diversité des appellations de ce test) et dont les affiliations sont aussi généralement bien renseignées et non normalisées.

Nous avons lancé des requêtes très larges sur ces deux sources, les plus inclusives possibles, afin d'obtenir un large éventail d'appellations.

[24] Wikipédia, novembre 2017

[25] Wikipédia, novembre 2017

Voici un exemple de la méthode de recherche employée pour l'Université de Lausanne sur HAL :

- Recherche avancée -> structure (multicritères) : Lausanne or Unil or chuv
- Tri : date de publication décroissante
- Seulement les articles parus dans une revue

Nous avons ensuite sélectionné les 100 premiers résultats qui concernent l'université testée dans chacune des deux archives pour obtenir un total de 200 notices.

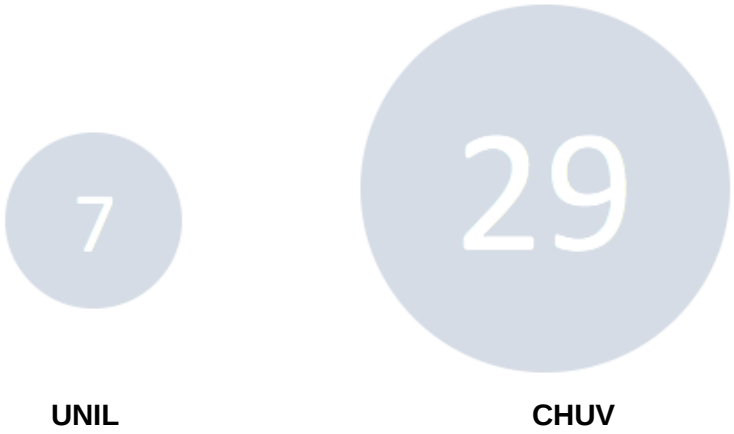
Nous avons listé pour chaque université toutes les appellations rencontrées dans un tableau Excel. Nous avions, à l'origine, conservé les appellations des facultés et des départements. Néanmoins, dans le traitement final des résultats, nous ne prenons en compte que les appellations globales des universités.

4.4.3 Analyse des appellations

Pour les deux universités testées nous avons choisi de séparer les résultats entre les appellations concernant les articles affiliés aux facultés internes à l'université et ceux affiliés à l'hôpital universitaire, en raison de la grande disparité des résultats entre ces deux cas.

Figure 20 : Nombre d'appellations différentes trouvées pour l'Université de Genève

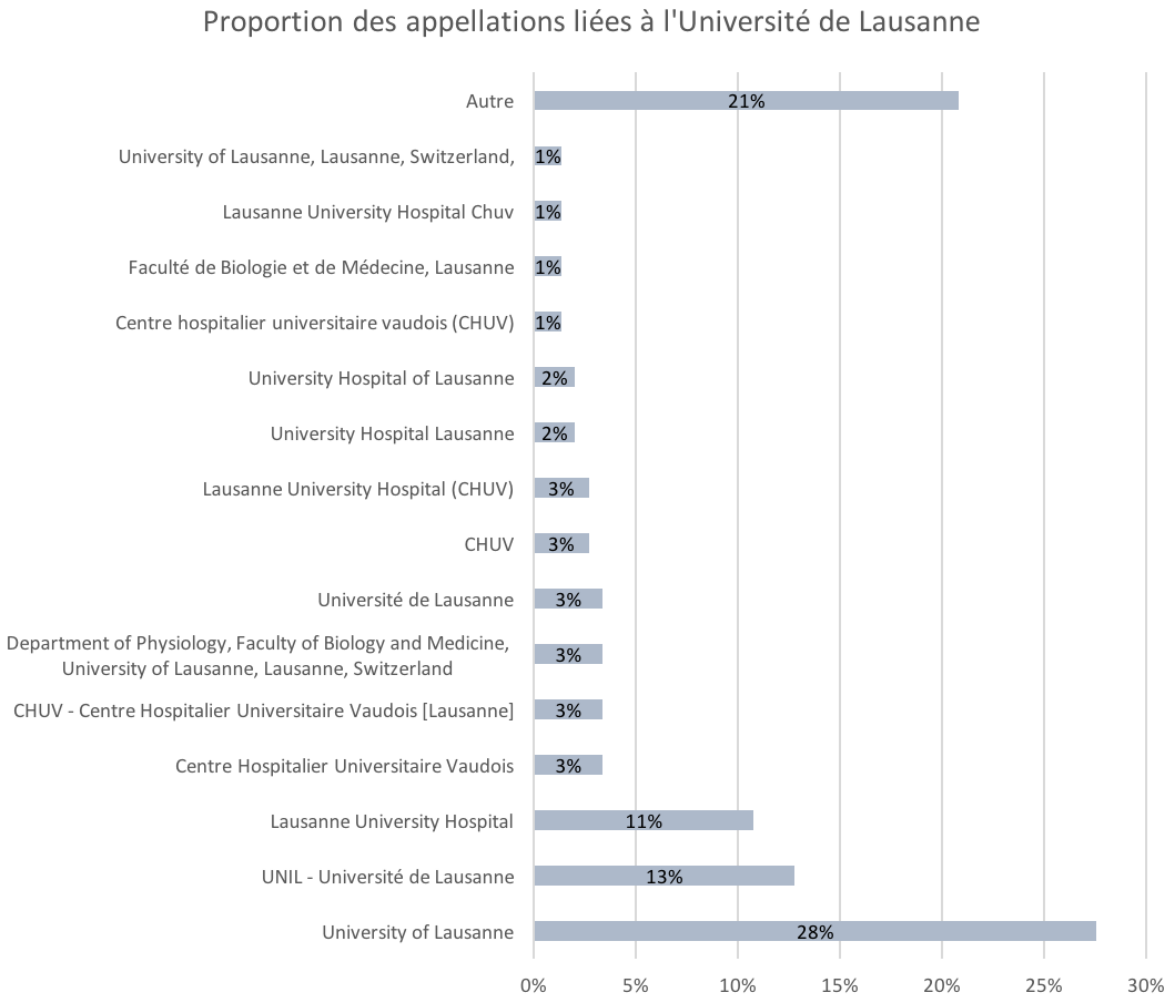
Figure 21 : Nombre d'appellations différentes trouvées pour l'Université de Lausanne



Les appellations varient énormément dans le cas des articles produits dans le milieu médical. Nous avons remarqué que les départements et laboratoires étaient souvent mis en avant par rapport à l'université.

27 et 29 appellations différentes pour un échantillon de 200 notices est un résultat très élevé, ce qui cause un problème d'envergure quant à l'identification des articles affiliés à une institution.

Figure 22 : Proportion des appellations utilisées pour nommer son affiliation à l'Université de Lausanne



D'après la directive de 2009 de l'Université de Lausanne concernant l'affiliation des chercheurs, il n'y a que quatre appellations retenues : deux pour l'hôpital et deux pour l'université en français et en anglais.

- Université de Lausanne (**3%** de notre échantillon) ;
- University of Lausanne (**28%**) ;
- Centre Hospitalier Universitaire Vaudois (**3%**) et Université de Lausanne ;
- Lausanne University Hospital (**11%**) and University of Lausanne.

En additionnant les différentes occurrences des appellations agréées dans notre échantillon nous arrivons à un total de **45%** d'appellations correctes.

Cela signifie que plus de la moitié des chercheurs ne suit pas la directive mise en place par l'université et utilise une appellation non standardisée. Nous constatons donc que la situation a peu évolué depuis l'enquête de 2005 (BADOR et LAFOUGE).

La recherche que nous avons effectuée présuppose que les appellations de ces deux universités comportent au moins un des termes utilisés dans notre recherche. Ce qui signifie que l'on est incapable d'évaluer le taux d'articles actuellement introuvables car ne répondant à aucun de nos critères de recherche ou dont l'affiliation n'est pas renseignée. L'étude de Bador et Lafouge (2005) estimait la part d'articles ne mentionnant pas du tout le nom de l'Université Claude Bernard Lyon 1 à 44%.

Dans le cadre du développement d'un script pour automatiser la collecte du plein-texte, cela signifie qu'il faut composer une requête qui soit à la fois inclusive pour atteindre toutes les appellations identifiées et assez exclusive pour ne pas admettre tous les articles des universités d'une même ville. La diversité des appellations des affiliations est donc une contrainte majeure du développement futur d'une solution automatisée.

4.5 Limites de la proposition de stratégie

Nous avons rencontré plusieurs difficultés et contraintes au cours de la modélisation de notre stratégie. Dans ce chapitre nous allons brièvement revenir sur les éléments les plus problématiques.

4.5.1 Appellations non normalisées

Comme vu précédemment, la diversité des appellations des affiliations est un enjeu majeur dans l'identification des articles.

Il s'agit de la première étape du processus : réussir à définir si un article a été écrit par un chercheur financé par une institution spécifique. Pour l'instant, nous n'avons pas trouvé de système d'identification efficace.

4.5.2 Problème des métadonnées

Les métadonnées d'affiliations ne sont souvent pas requises par les sources. Sur Crossref, une grande partie des articles n'a pas les champs "funders" ni "affiliation" renseignés, ce qui rend l'attribution d'un article à une institution très difficile.

Il s'agit également de la raison principale pour laquelle nous avons renoncé à utiliser Arxiv dans la première partie de notre recherche. En effet, on ne peut pas déterminer sans effectuer une recherche dans le plein-texte quelle en est l'affiliation. Nous avons envisagé de télécharger la totalité des plein-texte et d'exécuter un script allant rechercher l'affiliation directement dans ceux-ci néanmoins cette solution s'est avérée techniquement trop lourde en raison du nombre très élevé de PDF à télécharger et de la diversité des appellations qui rend difficile la mise en place d'une requête efficace.

4.5.3 Diversité des sources des PDF

En ce qui concerne l'utilisation d'Open Access Button, nous rencontrons deux problèmes fondamentaux. Le premier est la diversité des sources dans lesquelles rechercher le PDF. 52% des articles en plein-texte que l'on peut obtenir par OAB nous renvoie d'abord sur la page web depuis laquelle le PDF est accessible. Identifier la totalité des sources et la manière dont le PDF est identifiable par un script demanderait un très gros travail en amont même si l'on estime que 78% des articles sont disponibles à partir de deux sources différentes.

L'autre problème important que nous avons rencontré avec OAB est le nombre de liens erronés. En effet, nous sommes renvoyés de nombreuses fois sur des dépôts institutionnels (par exemple Infoscience) ou sur des plateformes d'éditeurs (Wiley) qui nous renvoient certes vers l'article en plein-texte, mais sur sa version payante qui est donc inaccessible.

5. Conclusion et recommandations

L'Open Access joue désormais un rôle central dans la diffusion et l'accès à la production académique. La progression de la part de marché de l'OA, bien que discrète, ne cesse de croître et est désormais encouragée et incitée dans toutes les universités suisses, notamment avec la mise en œuvre de la stratégie nationale suisse de l'Open Access du FNS depuis janvier 2017.

Pour les dépôts institutionnels, cela représente un potentiel de développement immense. En effet, dans un avenir marqué par l'avènement de l'OA, nous pourrions imaginer des dépôts institutionnels couvrant l'entier de la production documentaire académique de leur institution.

C'est pourquoi ils ont aujourd'hui l'obligation de se positionner en tant que faire-valoir du travail des chercheurs de leurs universités et de les convaincre de la nécessité d'y déposer leurs articles. Les dépôts doivent être en mesure de présenter une image fidèle et complète de la production académique de leurs institutions et de leur dynamisme afin d'en être une vitrine pour l'extérieur.

Dans ce projet de recherche nous avons voulu savoir où se situaient aujourd'hui les dépôts institutionnels suisses en matière d'articles en plein-texte afin d'en estimer le potentiel d'augmentation. Cela nous a permis de découvrir d'importantes variations entre les sept dépôts testés et de développer un modèle de stratégie d'automatisation de la collecte du plein-texte manquant.

Nous avons calculé un potentiel d'augmentation des dépôts institutionnels doublant leur couverture actuelle. Ils seraient en mesure de passer de 35,5% à 74,8% d'articles en plein-texte. Ce potentiel relativement élevé justifiant ainsi un travail de modélisation d'une stratégie d'automatisation de la collecte des plein-textes.

Il a néanmoins été difficile de modéliser une solution technique alors que la qualité des données ne suivait pas. En effet, nous avons été confrontés à divers problèmes liés à l'incomplétude des données disponibles. Les métadonnées liées au financement ou à l'affiliation étaient souvent manquantes ou incorrectes ce qui a rendu la détermination des affiliations des articles très ardue.

Nous avons toutefois développé, malgré les difficultés rencontrées, une stratégie d'automatisation basée sur Crossref et Open Access Button qui permet de rapporter en moyenne 20% des articles manquants.

5.1 Recommandations

5.1.1 Sensibiliser les chercheurs

Dans notre étude nous ne nous sommes pas concentrés sur les raisons qui font que les chercheurs ne déposent pas de manière systématique leurs articles sur le dépôt institutionnel de leur université. De nombreuses autres études se sont déjà consacrées à cette problématique mais il est vrai qu'en comprenant ce qui empêche les chercheurs de déposer leurs articles sur le dépôt, il est possible d'augmenter considérablement son taux de couverture.

Les chercheurs ont tendance à privilégier le dépôt de leurs articles dans les archives ouvertes liées à leur domaine plutôt que le dépôt de leur institution car cela leur offre une plus grande visibilité dans leur domaine d'études. (NARAYAN et LUCA 2016)

Or, le soutien des chercheurs est indispensable dans l'accroissement du fonds d'un dépôt institutionnel.

Dans leur étude de 2016 sur les difficultés et les défis de l'adoption d'un dépôt institutionnel par les chercheurs, Narayan et Luca ont relevé les barrières principales empêchant les chercheurs d'utiliser le dépôt :

- Des problèmes liés à l'outil même (navigation et architecture de l'information)
- Une méconnaissance de l'Open Access

Ainsi, une meilleure sensibilisation des chercheurs à l'Open Access ou une incitation à la publication sur ce modèle (comme cela se fait déjà au CHUV[26]) pourrait les encourager à utiliser plus systématiquement le dépôt et à reconsidérer leur approche de la publication académique traditionnelle. Une plus grande implication des chercheurs dans la conception du dépôt institutionnel en matière d'architecture de l'information permettrait une meilleure adéquation de l'outil avec leurs besoins et faciliterait donc la prise en main du dépôt.

5.1.2 ORCID

Depuis janvier 2017, ORCID[27] (organisation attribuant des identifiant uniques et pérennes aux chercheurs) a lancé un groupe de travail sur un projet d'identifiant unique institutionnel : ORGID. En s'associant, entre autres, avec Crossref, ils envisagent le développement d'un identifiant unique qui associe à la fois l'identifiant unique de l'auteur, celui de l'institution à laquelle il est affilié et celui de l'article.

Cela permettrait aux institutions de pouvoir facilement identifier les publications qui leur sont affiliées et réduirait ainsi nettement la difficulté actuelle liée aux appellations.

[26] https://www.bium.ch/news/subsides-open-access-chercheurs-de-fbm/

[27] https://orcid.org/

Il nous semble que la question des appellations reste un problème sur lequel travailler. Notre étude n'a pas trouvé de solution à la difficulté d'identification des articles publiés par des auteurs financés par la recherche suisse. De plus, nous ne nous sommes pas consacrés à l'estimation du nombre d'articles publiés par chercheurs affiliés à des institutions suisses qui ne mentionnent en rien leur affiliation. Il s'agit donc d'articles perdus, dont les fonds dépensés par l'institution ne lui seront pas rendus en termes de prestige.

Nous pensons que les dépôts institutionnels ont un rôle essentiel à jouer dans la valorisation du travail des chercheurs et la mise en avant du dynamisme d'une institution. Il est donc primordial de mettre le dépôt en avant au sein de l'institution.

6. Bibliographie

ALEXANDRIA, 2017, *Research Platform Alexandria* [en ligne], University of St.Gall, [Consulté le 17 mai 2017] disponible à l'adresse : https://www.alexandria.unisg.ch/cgi/stats/report

ARCHIVE OUVERTE HAL, 2017, *HAL Archive ouverte* [en ligne], [Consulté le 19 décembre 2017] disponible à l'adresse : https://hal.archives-ouvertes.fr/

ARCHIVE OUVERTE, 2017, *Archive Ouverte UNIGE* [en ligne], Université de Genève, [Consulté le 17 mai 2017] disponible à l'adresse : https://archive-ouverte.unige.ch/

ARODES OPEN ARCHIVE, 2017, *Arodes open archive* [en ligne], Hes-so, [Consulté le 17 mai 2017] disponible à l'adresse : http://arodes.hes-so.ch/

ARXIV, 2017, arXiv. Org [en ligne], 2017, Cornell University Library, [Consulté le 10 mai 2017] disponible à l'adresse : https://arxiv.org/

AUBRY, Christine, JANIK, Joanna, 2005, *Les archives ouvertes Enjeux et pratiques : Guide à l'usage des professionnels de l'information,* Paris *:* ADBS Ed., ISBN 2-84365-079.8

BADOR, Pascal, LAFOUGE, Thierry, 2005, Rédaction des adresses sur les publications : Un manque de rigueur défavorable aux universités françaises dans les classements internationaux, *La presse médicale,* 14 mai 2005, t. 34 pp. 633 - 636, ISSN : 0755-4982

BETHUNE, Jörn, KRAEMER, Lars, THOMSEN, Ingo, et al., 2017, LitDB : Keeping Track of Research Papers From Your Institute Made Simple, Source Code for Biology and Medicine, 21 mars 2017, 2017;12:5. doi:10.1186/s13029-017-0065-2

BIUM, 2017, Des subsides Open Access pour les chercheurs de la FBM, *Bibliothèque universitaire de médecine* [en ligne], CHUV, [Consulté le 15 janvier 2018] disponible à l'adresse : https://www.bium.ch/news/subsides-open-access-chercheurs-de-fbm/

BORIS, 2017, *Bern Open Repository and Information System* [en ligne], Université de Berne, [Consulté le 17 mai 2017] disponible à l'adresse : https://boris.unibe.ch/

BOUDRY, C, DURAND-BARTHEZ, M, 2017, Publications en libre accès en biologie—médecine : historique et état des lieux en 2016, *Ethics, Medicine and Public Health,* 2017, v. 3, pp. 169 - 181, ISSN : 2352-5525

COMMISSION RECOMMENDATION of 17.7.2012 on access to and preservation of scientific information, 2012, *ec.europa.eu* [en ligne], 17 juillet, 2012, [Consulté en octobre 2017] disponible à l'adresse : http://ec.europa.eu/research/science-society/document_library/pdf_06/recommendation-access-and-preservation-scientific-information_en.pdf

CROSSREF, 2017, *Crossref* [en ligne], [Consulté le 12 novembre 2017] disponible à l'adresse : https://www.Crossref.org/

DECLARATION DE BERLIN, 2003, [trad. française], *Déclaration de Berlin sur le Libre Accès à la Connaissance en sciences exactes, sciences de la vie, sciences humaines et sociales* [en ligne], [Consulté le 23 avril 2017] disponible à l'adresse : http://openaccess.inist.fr/?Declaration-de-Berlin-sur-le-Libre

DECLARATION DE BETHESDA, 2003, [trad. française], *Déclaration de Bethesda pour l'édition en libre accès* [en ligne], Summary of the April 11, 2003, Meeting on Open Access Publishing, [Consulté le 23 avril 2017] disponible à l'adresse : http://openaccess.inist.fr/?Declaration-de-Bethesda-pour-l#n1

DECLARATION DE BUDAPEST, 2002, [trad. française], *Initiative de Budapest pour l'Accès Ouvert* [en ligne], [Consulté le 23 avril 2017] disponible à l'adresse : http://openaccess.inist.fr/?Initiative-de-Budapest-pour-l

DOCUMENTATION FOR CROSSREF'S REST API, 2017, *GitHub* [en ligne], [Consulté en novembre 2017] disponible à l'adresse : https://github.com/Crossref/rest-api-doc#queries

DOI.ORG, 2017, Registration Agencies - Areas of Coverage, *doi.org* [en ligne], 7 novembre 2017, [Consulté en décembre 2017] disponible à l'adresse : https://www.doi.org/RA_Coverage.html

EDOC, 2017, *Edoc* [en ligne], Université de Bâle, [Consulté le 17 mai 2017] disponible à l'adresse : http://edoc.unibas.ch/

EUROPE PMC, 2017, *Europe pmc* [en ligne], [Consulté en décembre 2017] disponible à l'adresse : https://europepmc.org/

HAAK, Laure, 2016, Organization identifier project : A way forward, *orcid.org* [en ligne], 31 octobre 2016, [Consulté en novembre 2017] disponible à l'adresse : https://orcid.org/blog/2016/10/31/organization-identifier-project-way-forward

HAAK, Laure, 2017, Organization Identifier Working Group: Update, *orcid.org* [en ligne], 18 septembre 2017, [Consulté en novembre 2017] disponible à l'adresse : https://orcid.org/blog/2017/09/19/organization-identifier-working-group-update

HIMMELSTEIN, Daniel S, ROMERO, Ariel R, MCLAUGHLIN, Stephen R, et al., 2017, Sci-Hub provides access to nearly all scholarly literature, 12 octobre 2017, *PeerJ Preprints* [en ligne] 5 : e3100v2 https://doi.org/10.7287/peerj.preprints.3100v2, [Consulté en novembre 2017] disponible à l'adresse : https://peerj.com/preprints/3100/

INFOSCIENCE, 2017, *Infoscience* [en ligne], Ecole polytechnique fédérale de Lausanne, [Consulté le 17 mai 2017] disponible à l'adresse : https://infoscience.epfl.ch/

KIKKAWA, J, TAKAKU, M, YOSHIKANE F, 2016, DOI Links on Wikipedia, *Morishima A., Rauber A., Liew C. (eds) Digital Libraries: Knowledge, Information, and Data in an Open Access Society* [en ligne], vol. 10075, [Consulté en décembre 2018], DOI : https://doi.org/10.1007/978-3-319-49304-6_40 disponible à l'adresse : https://link.springer.com/chapter/10.1007/978-3-319-49304-6_40

MOREIRA, Miguel, La Lettre 2017-1, *Rero* [en ligne], Avril 2017, [Consulté en mai 2017] disponible à l'adresse : https://www.rero.ch/pdfview.php?section=lalettre&filename=LaLettre2017_01.pdf

NARAYAN, Bhuva, LUCA, Edward, 2017, Issues and challenges in researchers' adoption of open access and institutional repositories: a contextual study of a university repository, *IR : Information research [en ligne], Proceedings of RAILS, Research Applications, Information and Library Studies, Victoria University of Wellington, New Zealand, 6-8 December, 2016,* 4 décembre 2017, [Consulté le 15 décembre 2017] disponible à l'adresse : http://www.informationr.net/ir/22-4/rails/rails1608.html

NEEDHAM, P, STONE, 2012, G, *IRUS-UK: making scholarly statistics count,* UK repositories Insights, 25(3), 262–267, doi: 10.1629/2048-7754.25.3.262

OPEN ACCESS BUTTON, 2017, *Open access button* [en ligne], [Consulté en novembre 2017] disponible à l'adresse : https://openaccessbutton.org/

ORGANIZATION IDENTIFIERS IN ORCID, 2017, *orcid.org* [en ligne], [Consulté en novembre 2017] disponible à l'adresse : https://members.orcid.org/api/resources/orgids-in-orcid

PEET, Lisa, 2016, Sci-Hub Controversy Triggers Publishers' Critique of Librarian, *LibraryJournal* [en ligne], 25 août 2016, [Consulté en novembre 2017] disponible à l'adresse : http://lj.libraryjournal.com/2016/08/copyright/sci-hub-controversy-triggers-publishers-critique-of-librarian/

PLOS, 2017, *PLOS* [en ligne], [Consulté le 11 janvier 2018] disponible à l'adresse : https://www.plos.org/

PUB MED. CENTRAL, 2017, *PMC* [en ligne], National center for biotechnology information, [Consulté en septembre 2017] disponible à l'adresse : https://www.ncbi.nlm.nih.gov/

RERO DOC, 2017, *Rero doc : Bibliothèque numérique*[en ligne], Rero, [Consulté le 17 mai 2017] disponible à l'adresse : https://doc.rero.ch/

SCHOPFEL, Joachim, PROST, Hélène, 2010, *Développement et Usage des Archives Ouvertes en France. 2e partie : Usage* [en ligne], 17 octobre 2010, disponible à l'adresse : https://archivesic.ccsd.cnrs.fr/sic_00527043

SCHOPFEL, Joachim, PROST, Hélène, 2010, *Les statistiques d'utilisation d'archives ouvertes : Etat de l'art* [en ligne], 4 mai 2010, disponible à l'adresse : https://archivesic.ccsd.cnrs.fr/sic_00480538

SCIENCEDIRECT, 2017, *Sciencedirect* [en ligne], Elsevier, [Consulté en avril 2017] disponible à l'adresse : https://www.sciencedirect.com/

SERVAL, 2017, *Serval : serveur académique lausannois* [en ligne], Université de Lausanne, [Consulté le 17 mai 2017] disponible à l'adresse : https://serval.unil.ch/

SHERPA ROMEO, 2017, *Sherpa/Romeo : opening access to research* [en ligne], [Consulté en mai 2017] disponible à l'adresse : http://www.sherpa.ac.uk/romeo/search.php

SSOAR, 2017, *Social Science Open Access Repository* [en ligne], [Consulté le 10 mai 2017] disponible à l'adresse : http://www.ssoar.info/en/home.html

SSRN, 2017, *SSRN* [en ligne], 2017, Elsevier, [Consulté le 17 mai 2017] disponible à l'adresse : https://www.ssrn.com/en/

STRATEGIE NATIONALE SUISSE SUR L'OPEN ACCESS, 2017, *swissuniversities.ch* [en ligne], 31 janvier 2017, [Consulté en octobre 2017] disponible à l'adresse : https://www.swissuniversities.ch/fileadmin/swissuniversities/Dokumente/Hochschulpolitik/Open_Access/Open_Access_strategy_final_f.pdf

THE OAISTER DATABASE, 2017, *OAIster* [en ligne], OCLC, [Consulté le 11 janvier 2018] disponible à l'adresse : https://www.oclc.org/en/oaister.html

UNIVERSITE DE LAUSANNE, 2009, *Directive de la Direction 4.3. : Affiliation des chercheurs* [en ligne], 23 avril 2007, [Consulté en décembre 2017] disponible à l'adresse : http://www.unil.ch/files/live/sites/fbm/files/shared/recherche/Affiliations_chercheurs/directive_affiliation_chercheurs.pdf

ZORA, 2017, *Zurich Open Repository and Archive* [en ligne], Université de Zürich, [Consulté le 17 mai 2017] disponible à l'adresse : http://www.zora.uzh.ch/

Annexe 1 : Cas rencontrés lors de la collecte d'articles

N° du cas	Identifiant de la première notice (Z1...)	Description du cas
1	hB01	L'article est librement téléchargeable en plein texte sur le dépôt institutionnel.
2	hB02	L'article est publié dans un journal avec une licence "Green" qui permet le partage du texte intégral. Cependant le document et sa notice est absent du dépôt institutionnel.
3	hB06	L'article est en pré-print. Cependant le document et sa notice est absent du dépôt institutionnel alors que la licence le permet.
4	hB04	L'article est soumis à un embargo encore en vigueur. Cependant l'article est disponible intégralement sur le dépôt institutionnel.
5	hB20	L'article ne peut pas être disponible pour sa diffusion en dehors du journal qui l'édite. Cependant la version post-print ou pré-print est disponible sur le dépôt institutionnel. Le format PDF de publication n'est disponible qu'aux utilisateurs authentifié (qui font partie de l'institution) sur le dépôt.
6	hB26	L'article ne peut pas être disponible pour sa diffusion en dehors du journal qui l'édite. Ni la version post-print, ni le pré-print n'est disponible sur le dépôt institutionnel. Le format PDF de publication n'est disponible qu'aux "utilisateurs authentifié" sur le dépôt.
7	hB28	L'article est disponible en pré-print sur la source (HAL ou PMC) mais c'est sa version publiée qu'on retrouve sur le dépôt institutionnel.
8	hB28	L'article n'est disponible qu'aux utilisateurs authentifiés pour une période limitée conformément à la période d'embargo.
9	hB34	L'article n'est disponible qu'aux utilisateurs authentifiés.

10	hE10	La notice seule est disponible sur le dépôt.
11	hE13	La licence de l'article interdit sa diffusion. L'article et sa notice sont absents du dépôt institutionnel.
12	hS01	La licence de l'article est indéterminée. L'article et sa notice sont absents du dépôt institutionnel
13	hS02	Le pré-print est disponible sur le dépôt.
14	hA17	Pas d'information sur la licence de l'article. L'article est disponible sur le dépôt pour les utilisateurs authentifiés.
15	hI01	L'accès au texte intégral sur le dépôt se fait au travers d'un lien vers le site de l'éditeur.
16	hI10	L'accès au texte intégral sur le dépôt se fait, moyennant un paiement, au travers d'un lien vers le site de l'éditeur.
17	pL06	L'article n'est pas disponible sur le dépôt. La revue ne permet ni la publication du PDF officiel ni celui du post-print.
18	pZ20	L'article n'est pas disponible sur le dépôt. Aucune information n'est disponible sur le statut de la revue.

Catégorisation des cas:

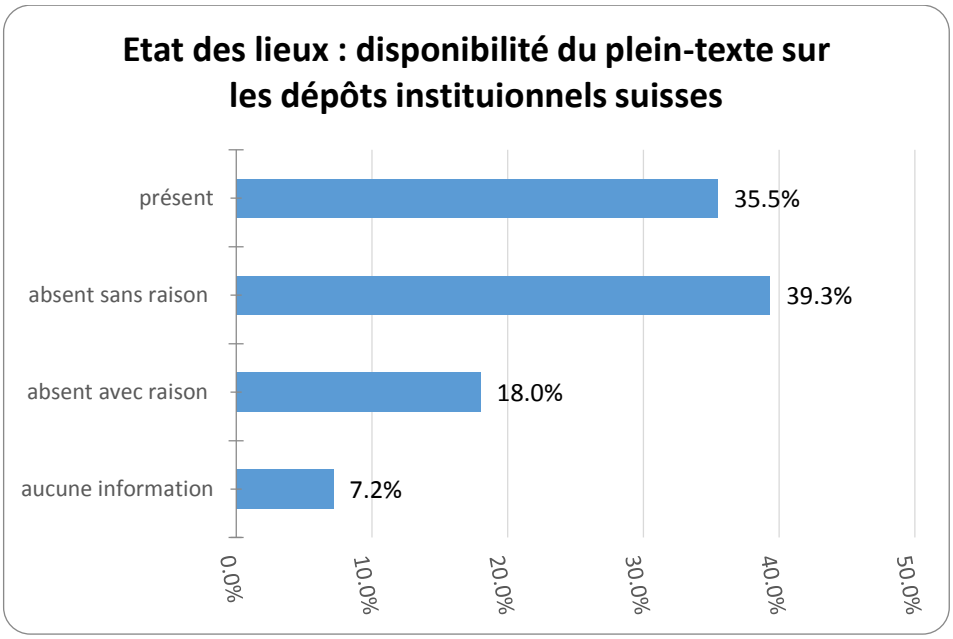
A. Présent : L'article est présent sur le dépôt institutionnel (cas 1 ; 4 ; 5 ; 7 ; 13)

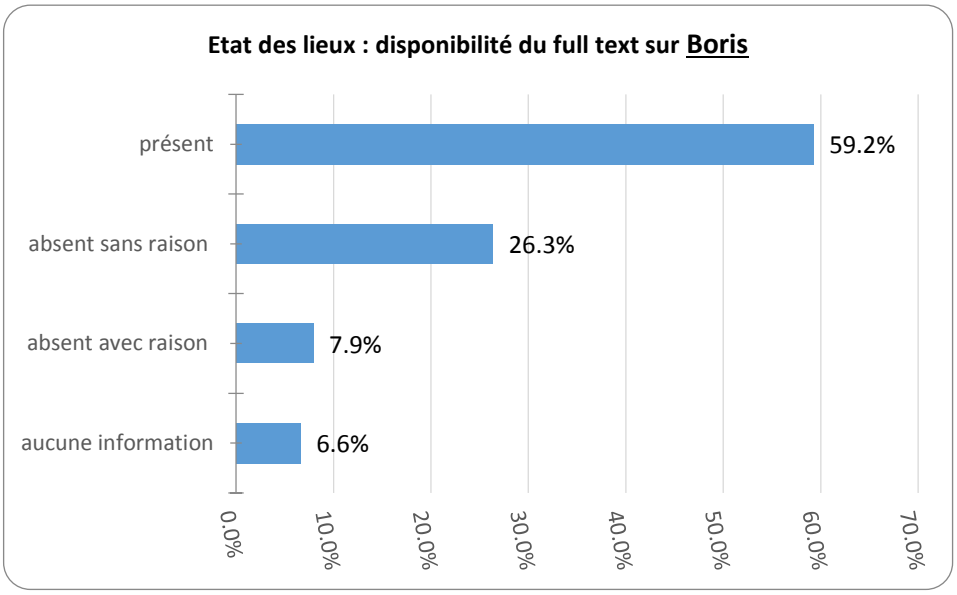
B. Absent sans raison. Il s'agit ici du potentiel d'augmentation des dépôts institutionnels (cas 2 ; 3 ; 6 ; 15).

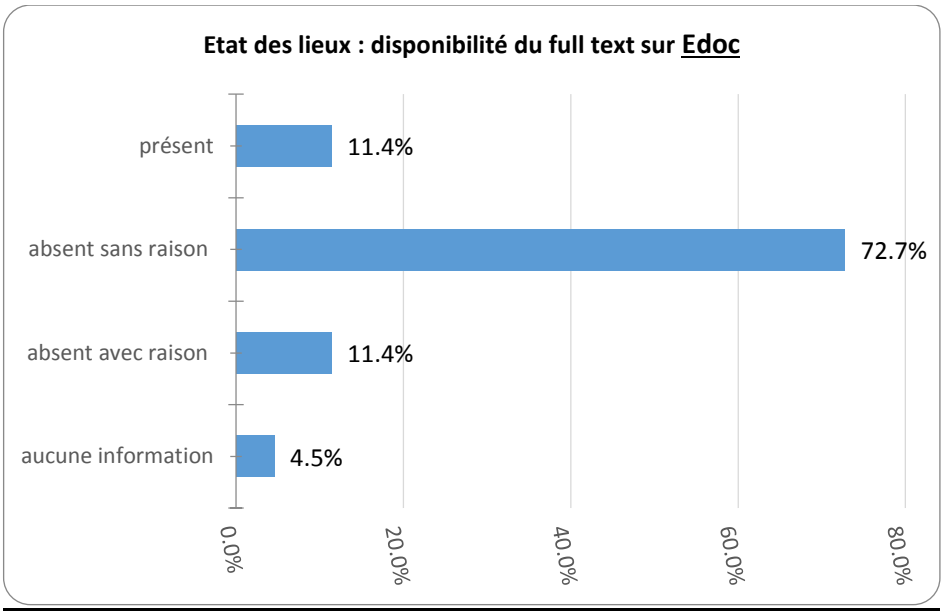
C. Absent : Absent du dépôt pour des questions de licences (cas 8 ; 9 ; 11 ; 16 ; 17)

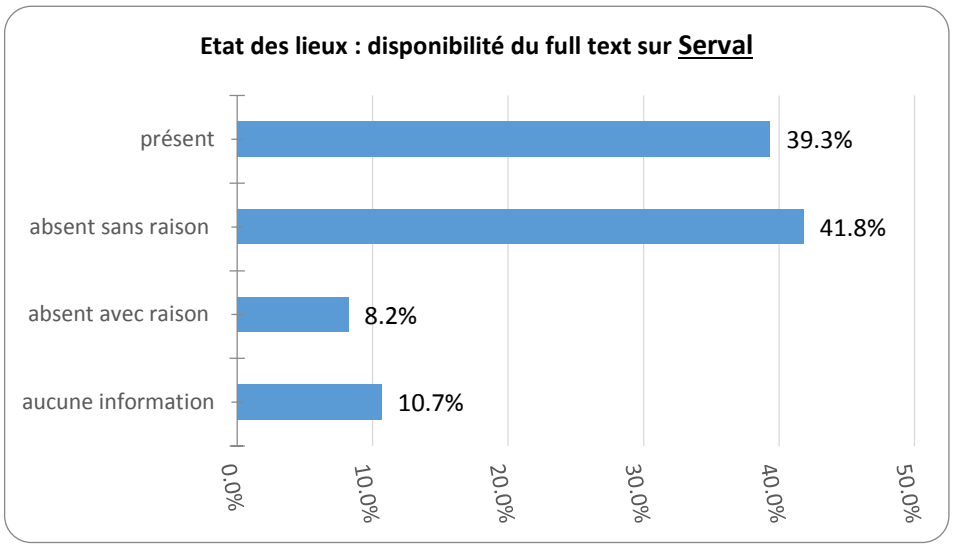
D. Autre : Pas de type de licence trouvé (cas 12 ; 14 ; 18)

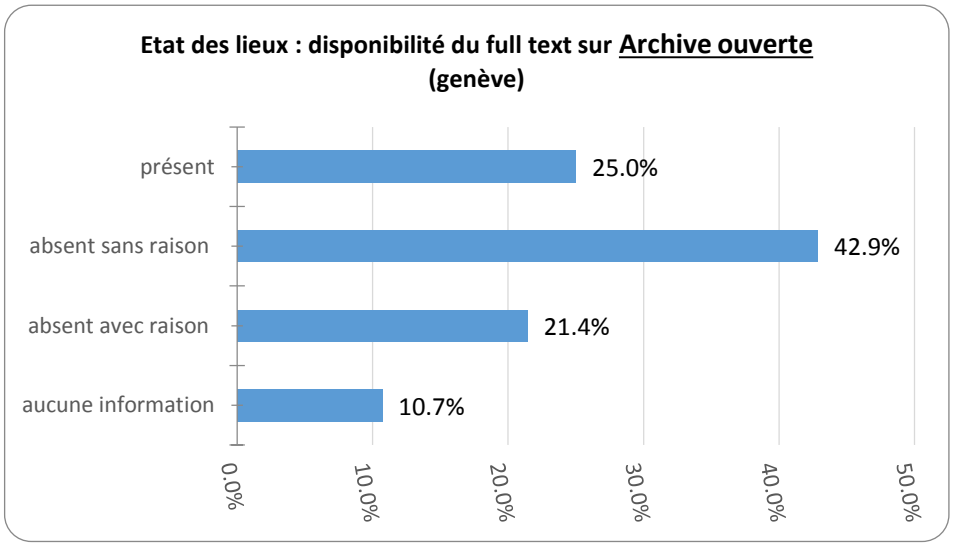
Annexe 2 : Graphiques de l'état des lieux des dépôts

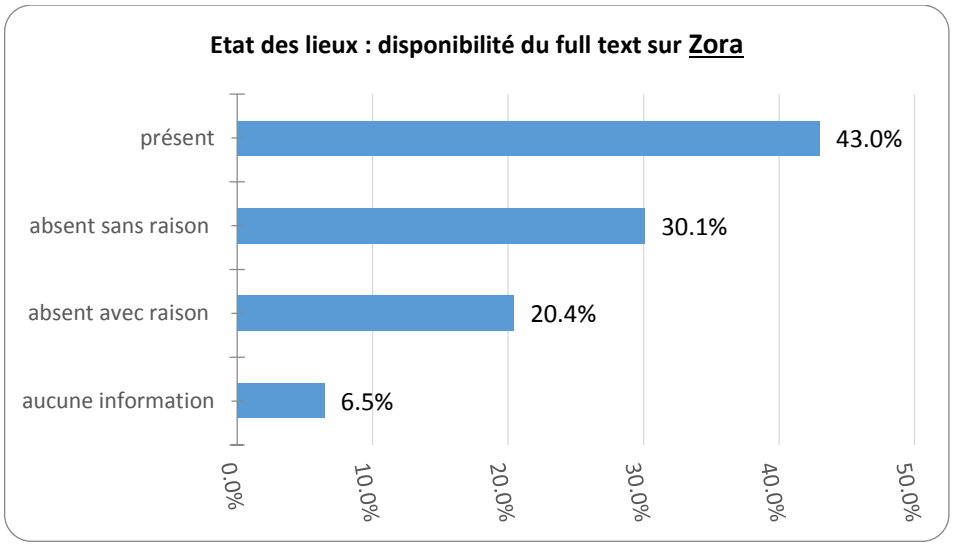


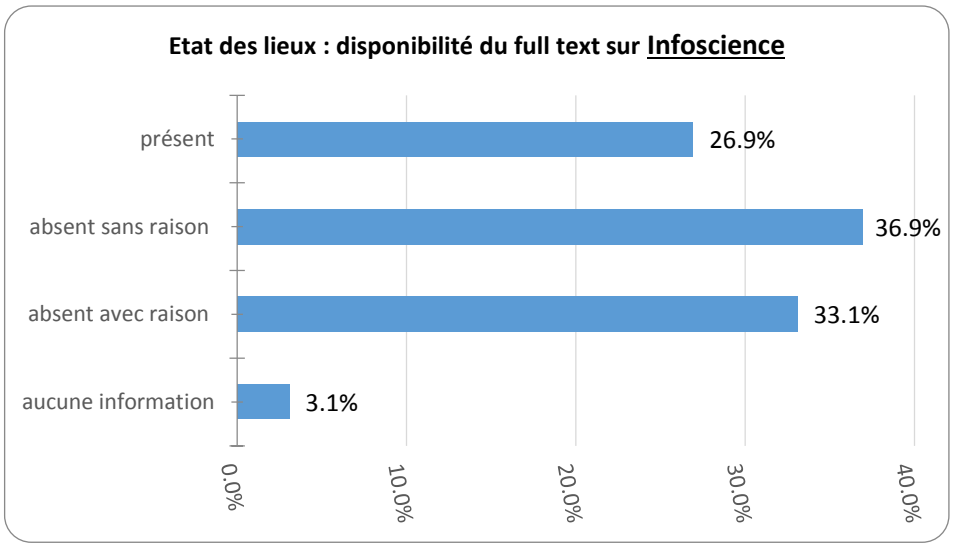


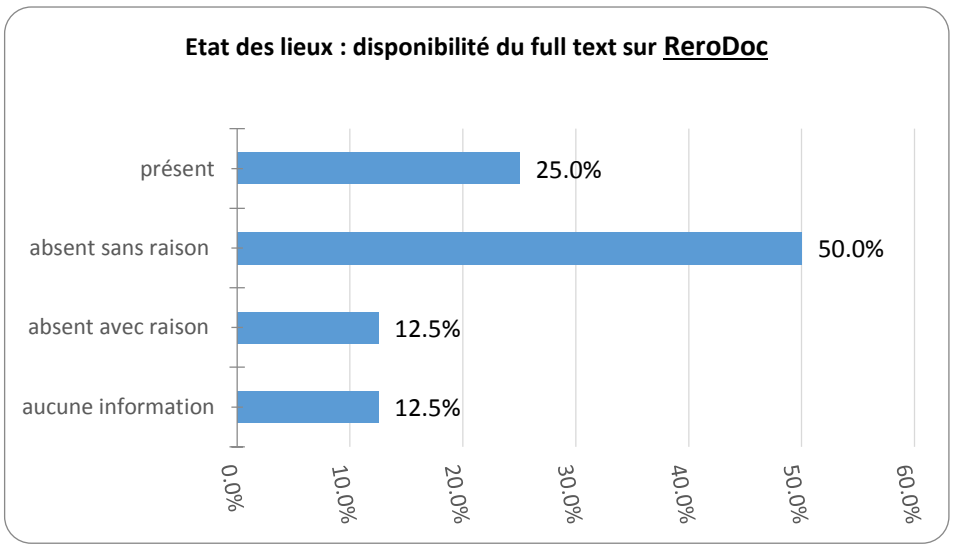

Etat des lieux : disponibilité du full text sur Edoc
présent
absent sans raison
absent avec raison
aucune information
11.4%
72.7%
11.4%
4.5%
0.0%
20.0%
40.0%
60.0%
80.0%


Etat des lieux : disponibilité du full text sur Serval
présent
absent sans raison
absent avec raison
aucune information
39.3%
41.8%
8.2%
10.7%
0.0%
10.0%
20.0%
30.0%
40.0%
50.0%


Etat des lieux : disponibilité du full text sur Archive ouverte (genève)
présent
absent sans raison
absent avec raison
aucune information
25.0%
42.9%
21.4%
10.7%
0.0%
10.0%
20.0%
30.0%
40.0%
50.0%


Etat des lieux : disponibilité du full text sur Zora
présent
absent sans raison
absent avec raison
aucune information
43.0%
30.1%
20.4%
6.5%
0.0%
10.0%
20.0%
30.0%
40.0%
50.0%


Etat des lieux : disponibilité du full text sur Infoscience
présent
absent sans raison
absent avec raison
aucune information
26.9%
36.9%
33.1%
3.1%
0.0%
10.0%
20.0%
30.0%
40.0%


Etat des lieux : disponibilité du full text sur ReroDoc
présent
absent sans raison
absent avec raison
aucune information
25.0%
50.0%
12.5%
12.5%
0.0%
10.0%
20.0%
30.0%
40.0%
50.0%
60.0%

Annexe 3 : Grille d'évaluation des outils de collecte du plein-texte

Nom	URL	Gouvernance	Type de solution	Domaines traités
Unpaywall	http://unpaywall.org/	Impactsory	Extension de navigateur	toutes disciplines
OpenAccessButt	https://openaccessbutton.org/	SPARC, a coalition of research libraries	Extension de navigateur et API	toutes disciplines
Core	https://core.ac.uk	Jisc, Open University	Agrégateurs d'articles en OA	toutes disciplines
SciHub	https://sci-hub.cc/	Alexandra Elbakyan	Base d'articles	toutes disciplines
Research gate	https://www.researchgate.net/		Réseau social scientifique	toutes disciplines
LibGen	http://gen.lib.rus.ec/	Communauté	Base d'articles	toutes disciplines
Freefullpdf	http://www.freefullpdf.com/index.html#gsc.tab=0	KnowMade SARL	Base d'articles	toutes disciplines
FindArticles	http://www.findarticles.com/	CBS Interactive Inc	Moteur de recherche	toutes disciplines
Open Edition	http://www.openedition.org/	Centre pour l'édition électronique ouverte		Sciences Humaines et sociales
AOIster	http://www.oclc.org/fr/oaister.html	OCLC	Catalogue collectif	toutes disciplines
Open Air	https://www.openaire.eu/	UE	Méta-catalogue	toutes disciplines

Nom	Sources des articles	API	Accès au fullte	Uniquement OA ou mixte	Recherche par DOI possib
Unpaywall	PubMed Central, the DOAJ, Crossref, DataCite, Google Scholar, and BASE	Oui	Oui	OA	avec l'API aoDOI
OpenAccessButt	oaDOI, Share, CORE, OpenAIRE, Dissem.in, Europe PMC, BASE (en gros tous les agrégateurs de IR)	Oui	Oui	Mixte	Oui
Core	6000 journaux et 3'659 IR	Oui	Oui	OA	Oui
SciHub	sites des éditeurs	Non	Oui	Mixte	Oui
Research gate			Oui	Mixte	Non
LibGen		Non	Oui	Mixte	Oui
Freefullpdf	26'000 sources		Oui	OA	Non
FindArticles			Oui	OA	Oui
Open Edition	464 revues		OUI	OA	Non
AOIster			Oui		
Open Air			Oui	Mixte	Non

Annexe 4 : Poster

ENRICHISSEMENT DES DÉPÔTS INSTITUTIONNELS SUISSES : VERS UNE COUVERTURE COMPLÈTE DE LA PUBLICATION ACADÉMIQUE OUVERTE

Stratégie d'automatisation du moissonnage de plein-textes

1

CONTEXTE

Les **dépôts institutionnels** pourraient être des acteurs clés dans la valorisation du travail des chercheurs et dans la visibilité de la **production académique** d'une université. Néanmoins, ces dépôts sont loin de l'exhaustivité et ne donnent accès qu'à une partie de cette production. Ce projet de recherche vise à quantifier la **proportion d'articles en plein-texte disponibles** sur ces dépôts et à développer une **stratégie de moissonnage automatique** des plein-textes afin d'**augmenter la couverture globale** des dépôts institutionnels suisses.

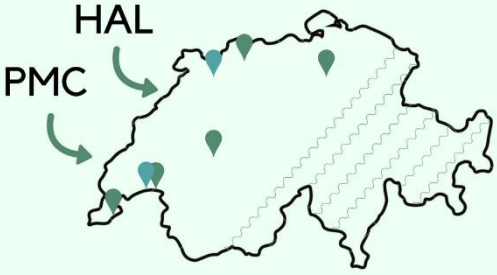
Dépôts institutionnels

Open Access

Publications

HAL

PMC



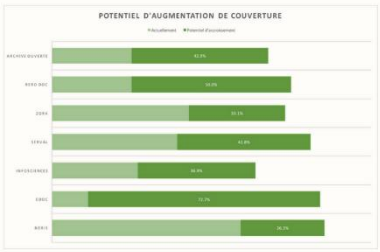
35.5%

des articles des chercheurs suisses sont disponibles en plein-texte dans les 7 dépôts testés.

2

MÉTHODOLOGIE

Nous avons sélectionné deux archives ouvertes d'envergure : l'une nationale et pluridisciplinaire **HAL**, l'autre internationale et monodisciplinaire **PMC**, dont nous avons comparé les résultats de requêtes spécifiques à ceux de **7 dépôts** institutionnels **Serval**, **InfoSciences**, **Boris**, **Zora**, **Edoc**, **Archive Ouverte** et **ReroDoc**. Pour les comparer, nous avons émis des requêtes dans HAL et PMC afin de trouver des articles en plein-texte parus en **2015** et **2016** affiliés à chaque institution suisse sélectionnée. Nous avons ensuite recherché chaque article trouvé précédemment dans le dépôt institutionnel correspondant afin de vérifier la présence du plein-texte dans les dépôts. Nous avons ainsi traité plus de **540 articles**.



74.8%

des articles (ou preprints) pourraient être disponibles légalement sur les dépôts testés.

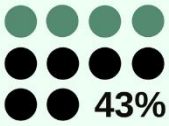
Nous avons identifié les articles qui n'étaient pas disponibles en plein-texte mais qui auraient logiquement dû l'être en raison de leur licence libre ou leur période d'embargo terminée. Ces articles représentent le **potentiel d'accroissement légal** des dépôts institutionnels testés.

3

VERS L'AUTOMATISATION

Nous avons émis une stratégie d'automatisation de la collecte du plein-texte qui se base sur l'extension de navigateur **Open Access Button** (OAB).
En partant de **140 DOI** d'articles affiliés à l'**UNIL** ou l'**UNIGE** dans **CrossRef**, nous avons vérifié la présence du plein-texte sur le dépôt, puis sa disponibilité avec OAB et avons finalement comparé avec la base de données controversée **SciHub**.
Nous avons testé manuellement notre stratégie qui s'est avérée n'apporter qu'un accroissement du fonds légalement disponible de 11% avec OAB.
Il s'agit là d'un accroissement en deçà de nos estimations, qui évaluaient le potentiel d'augmentation à environ 40% pour les 2 dépôts testés. SciHub permettrait, quant à lui, d'augmenter (sans tenir compte des licences) la couverture des dépôts à 96%.

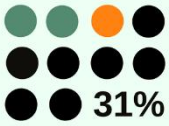
DÉPÔTS



43%

D'articles disponibles en plein-texte

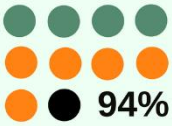
OAB



31%

Et un potentiel d'augmentation de 11%

SCIHUB



94%

Et un potentiel d'augmentation de 53%

1 UNIVERSITÉ
200 ARTICLES
36 APPELLATIONS

?

Combien d'articles introuvables à cause d'une appellation erronée?

4

LE PROBLÈME DES APPELLATIONS

Nous avons réalisé que le problème de fond se situait au niveau des **appellations des institutions**. En effet, malgré des efforts de **normalisation** des appellations, les institutions ont beaucoup de difficultés à imposer un modèle type à leurs chercheurs. Cela rend ainsi l'**identification** de nouvelles publications très ardue voire **impossible** dans le cas où l'Université n'est pas nommée dans la publication. Pour l'Université de Lausanne nous avons trouvé 36 appellations différentes sur un jeu de 200 notices.

Matthieu Putallaz et Elodie Schwob

matthieu.putallaz@etu.hesge.ch / elodie.schwob@hesge.ch
Sous la direction de Patrick Ruch, Professeur HES

Master en sciences de l'information, 2016-2018
Projet de recherche

Session Poster, le 14 décembre 2017

h e g
Haute école de gestion Genève