Évaluation de la production des laboratoires de recherche en SIC dans l'environnement de la science ouverte : analyse bibliométrique des publications sur HAL

Arezki Achouri

Arezki ACHOURI

Master Information Documentation
Première année

MÉMOIRE DE STAGE
Mission effectuée du 15 avril au 18 juin 2021

au
Laboratoire GERiiCO
Villeneuve d'Ascq

Évaluation de la production des laboratoires de recherche en SIC dans l'environnement de la science ouverte : analyse bibliométrique des publications sur HAL

Sous la direction de :
M. J. SCHÖPFEL (tuteur universitaire)
M. S. CHAUDIRON (tuteur professionnel)

Soutenu le 28 Juin 2021 à l'UFR DECCID-SID
Université de Lille, Sciences Sociales et Humaines
BP 60 149, 59 653 Villeneuve d'Ascq Cedex

Année universitaire 2020/2021

Évaluation de la production des laboratoires de recherche en SIC dans l'environnement de la science ouverte : analyse bibliométrique des publications sur HAL

Evaluation of the production of ICS research laboratories in the open science environment :

bibliometric analysis of publications on HAL

Remerciements

Je tiens en premier lieu à remercier très chaleureusement M. Joachim SCHÖPFEL et M. Stéphane CHAUDIRON de m'avoir fait confiance et sélectionné pour réaliser ce projet.

J'aimerais également leur adresser mes plus cordiaux remerciements pour leur suivi, leurs judicieux conseils, leur disponibilité ainsi que leur temps précieux qu'ils m'ont consacré tout au long des périodes de vacation et de stage.

Je tiens particulièrement à exprimer mes sincères remerciements et ma profonde gratitude à Mme Christelle BANTEGNIES (Gestionnaire pédagogique au département SID) pour ma convention et sa grande gentillesse. J'apprécie énormément ce qu'elle fait.

Je voudrais aussi remercier Mme Delphine SPILEERS (Responsable administrative et financière du GERiiCO) pour la gestion de mes dossiers de vacation et de stage.

Je souhaiterais enfin remercier ma famille et mes camarades : Amara COZZAROLO, Alexandre RIOTTE et Cerine BOUBEHA qui m'ont beaucoup soutenu et encouragé durant toute cette année.

Résumé :

Cette étude fournit une première approche d'évaluation de la production scientifique des laboratoires de recherche en Sciences de l'information et de la communication (SIC) sur la plateforme d'archives ouvertes HAL. Elle porte sur les publications de 36 laboratoires de recherche en SIC dans HAL et présente les résultats d'une analyse bibliométrique réalisée à partir des données extraites via HAL. L'objectif de cette étude est d'analyser la présence des laboratoires de recherche en SIC sur HAL en fonction d'un certain nombre de variables : le nombre de dépôts, la langue de publication, la typologie des documents et la part des documents en libre accès. Les résultats de l'étude ont montré une présence importante des laboratoires en SIC sur HAL, mais nous avons également constaté des différences entre les laboratoires de recherche concernant leur nombre dépôts, leur ouverture en termes de libre accès ainsi que l'internationalisation de leurs collections. Chaque laboratoire a sa propre politique de publication sur HAL.

Abstract :

This study provides a first approach to evaluate the scientific production of research laboratories in Information and Communication Sciences (ICS) on the open archive platform HAL. It focuses on the publications of 36 ICS research laboratories in HAL and presents the results of a bibliometric analysis realized from the data extracted via HAL. The objective of this study is to analyze the presence of ICS research laboratories on HAL according to a certain number of variables : the number of repositories, the language of publication, the typology of documents and the share of open access documents. The results of this study revealed an important presence of the laboratories in ICS on HAL, but we also observed differences between the research laboratories concerning their number of repositories, their opening in terms of open access and the internationalization of their collections. Each laboratory has its own publication policy on HAL.

Mots-clés :

Science ouverte, libre accès, archive ouverte, HAL, API, analyse bibliométrique, production scientifique, laboratoire de recherche en Sciences de l'information et de la communication.

Keywords :

Open science, open access, open archive, HAL, API, bibliometric analysis, scientific production, research laboratory in Information and Communication Sciences.

Table des matières

Introduction

Au cours de ces dernières années, la diffusion des données et des résultats de recherche dans l'environnement de la science ouverte, notamment en libre accès, s'impose peu à peu comme une nouvelle forme de publications scientifiques. De nombreuses plateformes d'archives ouvertes comme HAL (Hyper articles en ligne), des archives institutionnelles et des sites de dépôts ont été développées afin de valoriser la production scientifique et de permettre un accès ouvert aux publications scientifiques.

La plateforme HAL[1] a été créée en 2001 par le Centre pour la Communication Scientifique Directe[2] (CCSD) du Centre national de la recherche scientifique[3] (CNRS). Il s'agit d'une archive ouverte pluridisciplinaire ayant pour vocation de diffuser des publications de niveau recherche qu'elles publiées ou non, du secteur privé ou public. Depuis le développement de la plateforme HAL, des études ont été menées sur l'usage de cette plateforme par les établissements d'enseignement supérieur, par les chercheurs afin de comprendre leur attitude à l'égard de cette archive ouverte et de la mise à disposition de leurs productions scientifiques en libre accès. Cependant, aucune étude ne s'est intéressée aux laboratoires de recherche qui sont des acteurs majeurs de l'activité scientifique jusqu'à ce que le projet HAL/LO[4] voit le jour. Un projet mettant en avant la production des laboratoires de recherche et la place de leurs collections dans HAL. Il est financé par le GIS Réseau Urfist[5] et porté par une équipe constituée des chercheurs du laboratoire GERiiCO[6] de l'université de Lille, de l'URFIST de Bretagne et des Pays de La Loire[7] de l'université de Rennes 2, ainsi que du CCSD du CNRS.

Notre étude s'intègre donc au sein du projet HAL/LO, et s'intéresse à la production des laboratoires de recherche en Sciences de l'information et de la communication (SIC) sur HAL. Elle s'appuie sur une approche quantitative et s'inscrit dans une perspective bibliométrique afin d'analyser les collections HAL des laboratoires de recherche en SIC. Nous nous interrogeons sur la présence des laboratoires de recherche en SIC sur HAL et

1 HAL archive ouverte : https://hal.archives-ouvertes.fr/

2 CCSD : https://www.ccsd.cnrs.fr/

3 CNRS : http://www.cnrs.fr/

4 Projet HAL/LO : https://gis-reseau-urfist.fr/hal-lo-valorisation-sur-hal-de-la-production-des-laboratoires-dans-lenvironnement-de-la-science-ouverte/

5 GIS Réseau Urfist : https://gis-reseau-urfist.fr/

6 GERiiCO : https://geriico.univ-lille.fr/

7 URFIST de Bretagne et des Pays de La Loire : https://urfist.univ-rennes2.fr/

à partir de quel moment ont-ils commencé à déposer leurs productions sur HAL ? Quels sont les types des documents déposés dans les collections HAL des laboratoires en SIC ? Quelle est la part des documents en libre accès dans leurs collections HAL ? Et quelle est la place réservée aux publications en langues étrangères dans les collections de ces unités de recherche ? Ce sont les questions auxquelles nous tenterons de répondre dans le cadre de cette première approche de notre étude.

Pour ce faire, nous avons dans un premier temps constitué une liste de laboratoires de recherche en SIC. Puis, nous avons procédé à l'extraction des données à partir de HAL. Nous nous sommes donc appuyés sur ces données extraites via HAL pour étudier la production des laboratoires de recherche en SIC sur HAL.

Enfin, notre plan est structuré en trois sections : dans la première section, nous nous intéresserons au cadre théorique et conceptuel et ferons également un état de l'art des recherches en lien avec notre sujet. Dans la seconde section, nous détaillerons la méthodologie mise en place, les objectifs de l'analyse ainsi que les difficultés méthodologiques rencontrées. Dans la dernière section, nous présenterons et analyserons les résultats de notre recherche.

I. Cadre théorique et conceptuel : état de l'art

Dans cette première section, nous allons nous intéresser à la partie théorique et conceptuelle de notre sujet, car il nous semble qu'il est important de déterminer les théories et de définir les concepts en lien avec notre étude afin de démontrer les bases sur lesquelles repose notre recherche. Dans un premier temps, nous évoquerons quelques travaux préexistants et similaires à notre étude. Dans un second temps, nous définirons selon différentes théories les concepts de bibliométrie, scientométrie et infométrie. Dans un troisième temps, nous aborderons la question des frontières disciplinaires pour un laboratoire de recherche en Sciences de l'information et de la communication.

1 Description des travaux de même nature :

L'évolution du nombre des publications scientifiques dans l'environnement de la science ouverte, notamment dans l'archive ouverte HAL a conduit quelques chercheurs à mener des études autour de cette plateforme. En effet, certains chercheurs se sont intéressés à l'étude de la plateforme elle-même, c'est-à-dire à ses fonctionnalités, aux services qu'elle propose, etc. D'autres ont mené des études sur les pratiques et les usages de cette plateforme. Ainsi, des analyses quantitatives ont été réalisées sur les publications de HAL.

Parmi ces quelques recherches qui existent et qui sont en lien avec le cadre de notre étude, nous pouvons en citer certains travaux de Joachim Schöpfel, ceux d'Annaïg Mahé et Camille Prime-Claverie ainsi que le travail d'Emile Gayoso.

Tout d'abord, Joachim Schöpfel, maître de conférences en Sciences de l'information et de la communication à l'Université de Lille et membre du laboratoire GERiiCO, a réalisé une étude sur l'usage de la plateforme HAL par les unités de recherche de l'Université de Lille[8]. Il s'agit donc d'une étude menée sur soixante-deux laboratoires de recherche de différents domaines et disciplines appartenant à l'Université de Lille. L'étude visait à savoir comment les laboratoires de l'Université de Lille utilisent la plateforme HAL et quelles stratégies qui ont été mises en place par ces laboratoires de

8 Schöpfel Joachim, « *L'usage de la plateforme HAL par des unités de recherche. Le cas de l'Université de Lille* », I2D - Information, données & documents, 2020/3 (n° 3), p. 167-198. DOI : 10.3917/i2d.203.0167. URL : https://www.cairn.info/revue-i2d-information-donnees-et-documents-2020-3-page-167.htm

recherche pour engager les chercheurs à déposer leurs publications dans la plateforme HAL. Elle s'est, en effet, intéressée à la présence des laboratoires de l'Université de Lille sur HAL au moyen d'une collection, au nombre de leurs publications et à la part des documents en libre accès, tout en distinguant les différences disciplinaires de ces unités.

Le panel des soixante-deux unités était composé essentiellement des unités mixtes (52 %), des unités propres de l'Université de Lille (37 %) mais aussi des unités de recherche relevant de l'Inserm avec seulement 7 %. Selon Joachim Schöpfel, les laboratoires des domaines sciences et technologies et des sciences de la vie et de la santé représentent plus de deux tiers des laboratoires de Lille et que ceux des sciences humaines et sociales, droit, économie et gestion représentent seulement un tiers ds laboratoires[9].

Joachim Schöpfel a montré que l'ensemble des unités de recherche de l'Université de Lille sont identifiables sur HAL, mais seulement les unités de domaines des sciences humaines et sociales qui ont une collection sur HAL, alors que les unités des sciences de la vie et de la santé n'ont pas du tout et que les autres domaines ont plus au moins une collection sur HAL. Selon lui, 31 % (19 unités) des laboratoires avaient pour objectif de rendre visible leur production scientifique à travers leur collection HAL.

L'étude révèle également que les laboratoires de l'Université de Lille comptent plus de 50 000 dépôts en avril 2020 et que la moyenne des dépôts par laboratoire est de 811 dépôts, sauf que cette moyenne n'est pas vraiment représentative, car selon Joachim Schöpfel, la moitié des laboratoires ne dépassent pas 300 publications sur HAL, tandis que certains laboratoires comptent plusieurs milliers de publications sur HAL. L'auteur indique aussi que parmi les 50 266 dépôts des unités de Lille sur HAL, 63 % de ces dépôts sont effectués à l'intérieur d'une collection, et rappelle aussi qu'il n'y a que 31 % des unités qui représentent ces collections.

Concernant la part des documents en accès libre, Joachim Schöpfel a montré que 23 % des dépôts des unités de recherche de Lille sont effectués avec un fichier (texte intégral), donc plus de 11 500 documents en libre accès. Selon lui, ce pourcentage est quasiment le même, que celui de l'ensemble des dépôts en texte intégral sur HAL[10]. Ainsi, il souligne que les unités des sciences de la vie et de la santé sont plus ouvertes en terme de dépôts en texte intégral que les unités des autres domaines. Selon lui, leur taux d'ouverture s'élève à 52 %, puis 23 % pour les unités de sciences technologies, 20 % pour les sciences humaines et sociales et enfin 17 % pour les unités de droit, économie et gestion.

Selon les conclusions de Joachim Schöpfel, il y a une vaste variété d'approches et de pratiques relativement à la mise en place d'une collection sur HAL, au nombre de

9 Schöpfel Joachim, « *L'usage de la plateforme HAL par des unités de recherche. Le cas de l'Université de Lille* », 2020, p. 15.

10 Selon Joachim Schöpfel, 25 % des dépôts sur HAL sont accompagnés du fichiers.

dépôts ainsi qu'à la part des documents en libre accès. Son analyse a révélé plusieurs variations disciplinaires et des approches analogiques entre les unités de recherche. Enfin, Joachim Schöpfel est très passionné par la science ouverte. Ces travaux sont multiples et mène encore de nombreuses études dans le cadre de la science ouverte : libre accès, ouverture des données de recherche, etc. Il est notamment à la tête du projet de recherche HAL/LO[11] portant sur la valorisation de la production scientifique des laboratoires sur la plateforme HAL.

Ensuite, Annaïg Mahé[12] et Camille Prime-Claverie[13] ont conduit ensemble une étude portant sur les pratiques de dépôts sur HAL. Selon les auteures, l'objectif de cette étude est d'identifier et d'évaluer la contribution des différentes catégories d'acteurs ; de définir les logiques de dépôts ainsi que de comparer les logiques disciplinaires et/ou d'acteurs. Leur étude s'est déroulée en trois temps.

En effet, elles se sont intéressées dans un premier temps aux pratiques de dépôts dans le domaine des sciences de la vie (SDV). Cette première étape s'est déroulée en 2012-2013 et s'est basée sur les dépôts effectués sur une période de dix ans (entre 2002 & 2012) par les sciences de la vie. Dans un second temps, Annaïg Mahé et Camille Prime-Claverie ont étudié les pratiques de dépôts en sciences humaines et sociales (SHS) pour la période de 2002 et 2016. Enfin, la dernière phase a porté cette fois-ci sur les pratiques des déposants (contributeurs) en SHS. Ces deux dernières phases se sont déroulées en même période, 2016-2017.

L'étude s'est appuyée sur des données extraites sous format XML-TEI à partir du site de dépôt HAL en utilisant le protocole OAI-PMH (*Open Archives Initiative Protocol for Metadata Harvesting*). Le corpus est constitué uniquement des notices dont le volume est de 58 692 notices pour les sciences de la vie et de 336 160 notices pour les sciences humaines et sociales.

Pour la partie portant sur les sciences de la vie, les auteures ont observé qu'il y a plus de notices que de texte intégral, respectivement 76 % et 24 %. Elles ont également montré que les dépôts sont globalement effectués par des intermédiaires de l'institution (71 %), 26 % des dépôts sont réalisés par des auteurs et 3 % des dépôts sont faits par des intermédiaires externes. On voit donc qu'il y a peu de dépôts par les auteurs eux-mêmes, ce qui signifie que la logique communication scientifique directe n'est pas vraiment appliquée dans le domaine des sciences de la vie. Concernant les intermédiaires de l'institution, Annaïg Mahé et Camille Prime-Claverie ont indiqué qu'il y a plusieurs irrégularité au niveau de tous les types dépôts. Selon elles, cela pourrait s'expliquer par divers motifs : « *Il pourrait s'agir d'une mise à jour du recensement de la production scientifique dans le cadre du fait de campagnes d'évaluation (tous les quatre ans environ)*

11 GIS Réseau Urfist « *HAL/LO – Valorisation Sur HAL de La Production Des Laboratoires Dans l'environnement de La Science Ouverte* ». Consulté 1 juin 2021 https://gis-reseau-urfist.fr/hal-lo-valorisation-sur-hal-de-la-production-des-laboratoires-dans-lenvironnement-de-la-science-ouverte/

12 Maître de conférences à l'Urfist de Paris - École Nationale des Chartes.

13 Maître de conférences à l'Université Paris Nanterre.

ou d'un « effet de rattrapage » par le dépôt massif d'enregistrements de nouveaux entrants (institutions qui ouvrent un site de dépôt)[14] ».

En ce qui concerne les résultats de l'étude portant sur les sciences humaines et sociales, les auteures ont aussi observé que – selon quatre logiques de dépôts qu'elles ont déterminé au préalable[15] – le taux de la communication scientifique directe est très faible (seulement 12,1 % des dépôts). Selon Joachim Schöpfel, Annaïg Mahé et Camille Prime-Claverie ont distingué trois catégories de chercheurs en sciences humaines et sociales, avec quelques différences disciplinaires. Elles ont montré en effet qu'il y a des chercheurs qui déposent très fréquemment sur HAL, des chercheurs qui déposent de façon régulière ainsi que d'autres chercheurs qui déposent rarement, de façon irrégulière.

Enfin, l'étude d'Emile Gayoso[16] a, quant à elle, porté sur la diffusion des articles de recherche des revues des sciences humaines et sociales sur HAL et sur les réseaux sociaux académiques : Academia et ResearchGate. Son étude s'est basée sur une liste exhaustive de revues en SHS de différents domaines disciplinaires dont le nombre de cette liste est de 368 revues. L'étude s'est intéressée aux articles parus entre 2010 et 2018 dans la liste des revues retenues.

La collecte des données s'est faite en trois façons différentes : obtention directe des publications des revues publiées sur la plateforme CAIRN, extraction des données à partir de l'entrepôt OAI-PMH pour les revues diffusées par OpenEdition et un travail manuel qui consistait à établir une liste des articles des revues publiées sur d'autres plateformes, sur des sites propres ou uniquement en version papier. Pour les l'exaction des données de dépôts sur HAL, Emile Gayoso s'est appuyé sur l'outil de recherche avancée de HAL pour extraire les données des revues publiant sur HAL en utilisant soit le nom de la revue ou son identifiant ISSN ou e-ISSN. Ainsi, pour les l'extraction des données à partir des réseaux sociaux académiques : Academia et ResearchGate, il a utilisé un script permettant d'automatiser des requêtes dans les bases de données de ces réseaux.

Les résultats de l'étude d'Emile Gayoso montrent qu'il y a plus d'articles disponibles en texte intégral sur les réseaux sociaux académiques, notamment Academia, que sur HAL. Cela explique que les chercheurs des sciences humaines et sociales, avec des variations disciplinaires, privilégient plus le réseau Academia pour diffuser leurs publications que la plateforme HAL[17].

---

14 Prime-Claverie Camille, Mahé Annaïg, « *Sites de dépôt en libre accès et formes de médiations : quelles évolutions ?* », 2013, p. 137.

15 *Les logiques de dépôt sont déterminées à partir : du délai entre la date de création de la notice et la date de publication du document décrit ; de la présence ou non d'un fichier accompagnant la notice (texte intégral pour les documents écrits)*. (Annaïg Mahé et Camille Prime-Claverie, « *Qui dépose quoi sur Hal-SHS ? Pratiques de dépôts en libre accès en sciences humaines et sociales* », 2017, p. 4.

16 Ancien postdoctorant au Dicen-IdF, CNAM.

17 Emile Gayoso, « *La diffusion sur Hal, Academia et ResearchGate des articles de recherche des revues françaises de SHS* », 2020, p. 27-28.

En résumé, toutes ces études sont intéressantes et s'inscrivent dans le périmètre de la science ouverte et du libre accès. Leur point commun est qu'elles se sont intéressées à l'analyse des dépôts dans différents domaines et disciplines sur HAL. Des études basées sur des approches quantitatives qui présentent quelques similitudes à l'approche que nous optons pour notre étude. Cependant, notre analyse est très spécifique et vise uniquement le domaine des Sciences de l'information et de la communication. Ainsi, notre approche est à la fois quantitative et qualitative (cette partie ne sera pas traitée dans le cadre de ce mémoire.) et s'inscrit dans une perspective bibliométrique et scientométrique. C'est pourquoi nous aborderons les notions de bibliométrie, scientométrie et infométrie dans la partie suivante.

2 Bibliométrie, scientométrie et infométrie :

La bibliométrie est une méta-science permettant d'analyser l'activité et les réseaux scientifiques. Il s'agit d'un domaine ayant pour objet d'étude la science. L'analyse bibliométrique repose sur une approche quantitative et l'usage d'outils mathématiques et statistiques. Elle s'appuie donc sur des données quantitatives que nous pouvons recueillir par le comptage des publications d'une revue, d'une plateforme de dépôts scientifiques, etc. ou par l'extraction de certains éléments à partir des publications elles-mêmes tels que les citations.

Tout d'abord, nous considérons que l'origine du terme « bibliométrie » remonte aux années 1930. En effet, ce mot a été introduit par Paul Otlet en 1934 dans son ouvrage majeur intitulé « *Traité de la documentation – le livre sur le livre – théorie et pratique* ». C'est dans ce traité que Paul Otlet a développé les premiers principes de la bibliométrie. Il a fortement mis l'accent sur la nécessite de passer d'une évaluation qualitative à une étude quantitative du livre. Il a notamment écrit que « *Les sciences du livre (…) doivent tendre maintenant à introduire la mesure dans leurs investigations.*[18] » Pour lui, la mesure du livre est essentielle pour l'organisation des connaissances. Il considère en effet que la mesure comme une « *forme supérieure que prend la connaissance.*[19] », car elle permet de parvenir à une analyse objective des phénomènes. Selon P. Otlet, la bibliométrie serait une branche de la bibliologie qui prend en charge la mesure des livres : « *La bibliométrie sera la partie définie de la bibliologie qui s'occupe de la mesure ou quantité appliquée aux livres (arithmétique ou mathématique bibliologique).*[20] » Par la suite, Otlet propose également un ensemble d'indices de mesure. Il suggère notamment la stylistique, la stichométrie, la mesure des incunables et de la lecture, la bibliosociométrie et les coefficients. Pour Otlet, la bibliométrie résume les statistiques et permet de donner des indices de comparaisons. Mais il souligne également que « *la statistique du livre se*

18 P. Otlet, *Traité de la documentation*,1934, p. 14.

19 P. Otlet, *Traité de la documentation*, 1934, p. 13.

20 P. Otlet, *Traité de la documentation*, 1934, p. 14.

confond avec la bibliométrie.[21] » Car, selon lui, la statistique s'intéresse également aux tirages, à la circulation du livre, aux bibliothèques, etc. Enfin, Otlet termine la partie consacrée à la bibliométrie dans son « *Traité de la documentation* » par la notion de la « mathé-bibliologie » qui se rattache, selon lui, à tout ce qui est de la mesure du livre (statistiques du livre, bibliométrie). Il souligne également que la mathé-bibliologie constitue un langage permettant d'exprimer les rapports logiques entre les faits[22].

Le terme « bibliométrie » semble laisser de côté jusqu'à la fin des années 1960 lorsqu'un auteur anglo-saxon, Alan Pritchard, publie en 1969 un article intitulé « *Statistical bibliography on bibliometrics* » et dans lequel il reprend ce terme créé initialement par Paul Otlet pour désigner le volet métrique de la bibliologie. Alan Pritchard pense qu'il était nécessaire de redéfinir le domaine d'analyse quantitative des publications scientifiques dont les premiers travaux dans ce domaine remontent au début du XX^{e} siècle. Un domaine qui était couvert par l'expression « *statistical bibliography* » (Hulme, 1923) Il propose alors le terme « bibliométrie » pour mieux cerner ces travaux. Cela fait donc naître un nouveau domaine de recherches quantitatives, le domaine de la bibliométrie. Selon Pritchard, la bibliométrie renvoie à « *l'application des mathématiques et des méthodes statistiques aux livres, articles et autres moyens de communication.*[23] » Cela signifie que la bibliométrie est un ensemble de procédures permettant l'évaluation de la production scientifique des chercheurs à partir du nombre de leurs publications ainsi que des citations auxquelles leurs publications ont donné lieu. Nous pouvons donc dire que l'analyse bibliométrique se donne pour objet de mesurer les publications de la recherche scientifique que ce soit des ouvrages, des articles ou tout autre moyen de communication. Selon Jean-Max Noyer, les outils statistiques utilisés par la bibliométrie visent à élaborer des indicateurs concernant les publications des diverses pratiques de recherche[24]. Nous pouvons également souligner que l'approche bibliométrique permet de décrire et d'analyser la science par le biais de ses résultats et se fonde sur la conception que le principe de la recherche scientifique est la production des connaissances.

Selon Fidelia Ibekwe-SanJuan, la bibliométrie renvoie à l'étude de production des publications scientifiques que l'on considère comme témoins tangibles de l'activité de recherche scientifique[25]. La bibliométrie est une science empirique et repose sur des lois et théories. Les premières lois et théories bibliométriques remontent aux années 1920 et 1930. Parmi les précurseurs de l'approche bibliométrique, nous retrouvons Alfred James Lotka qui a mené une étude sur la répartition des auteurs en fonction du nombre des publications qu'ils produisent dans un domaine. En 1926, il établit une loi qui porte son nom : « la loi Lotka ». Cette dernière permet d'analyser la production des auteurs. Nous avons ensuite Samuel Clement Bradford qui a contribué aussi aux lois bibliométriques. En 1934, il met en place la « loi Bradford » lorsqu'il voulait étudier la diffusion des publications

21 P. Otlet, *Traité de la documentation*, 1934, p. 16.

22 P. Otlet, *Traité de la documentation*, 1934, p. 22.

23 Jean-Max Noyer, *Les sciences de l'information – Bibliométrie, Scientométrie, Infométrie*, 1995, p. 16.

24 Jean-Max Noyer, *Les sciences de l'information – Bibliométrie, Scientométrie, Infométrie*, 1995, p. 175.

25 Fidelia Ibekwe-SanJuan, *La science de l'information : origines, théories et paradigmes*, 2012, p. 177.

dans les revues scientifiques d'un domaine donné. Cette loi Bradford permet donc d'analyser les revues scientifiques afin de les modéliser. Nous retrouvons également George Kingsley Zipf qui a mis en place en 1935 la « loi Zipf ». Une loi permettant d'étudier la fréquence des mots dans un texte ou un corpus. Selon Fidelia Ibekwe-SanJuan, ces trois lois ne s'intéressent qu'aux publications de la recherche scientifique et qu'elles reposent toutes sur le même principe et parviennent à une même conclusion.

Ensuite, la notion de scientométrie apparaît dans les années 1960 qui se situe au prolongement de la bibliométrie. Elle est donc en partie liée avec la bibliométrie. Selon Jean-Max Noyer, le mot « scientométrie » serait synonyme des « études quantitatives de la science » et que les deux notions incluent un aspect bibliométrique. La scientométrie est apparue pour distinguer l'utilisation de la bibliométrie dans la gestion des bibliothèques et dans la circulation des publications scientifiques. Pour Michel Callon et al., la scientométrie est le nom donné à un ensemble de travaux consacrés à l'étude qualitative de l'activité de recherche scientifique et technique. Selon lui, l'approche scientométrique aurait pour but d'étudier aussi bien les ressources et les résultats que les formes d'organisation de la production des connaissances et de savoir faire[26].

Derek de Solla Price que l'on considère comme le père d'une « *science de la science* » a défini la scientométrie comme « *les recherches quantitatives de toutes les choses concernant la science et auxquelles on peut attacher des nombres.*[27] » Nous pensons que cette définition, donnée par Price à la scientométrie, est large, car elle pourrait réduire le sens de la scientométrie au même sens que celui de la bibliométrie. Selon Jean-Max Noyer, la scientométrie désigne « *l'application des méthodes statistiques à des données quantitatives caractéristiques de l'état de la science*[28] ». Pour Jean-Max Noyer, le domaine de la scientométrie a été développé pour répondre aux besoins de la politique de la science et de la gestion de recherche. Il est aussi le résultat d'usage des techniques statistiques et informatiques dans les études de la science. Ce domaine scientométrique recouvre plusieurs sous-domaines tels que le développement des indicateurs visant à mesurer les performances de la recherche et les performances technologiques ainsi que le développement des domaines scientifiques et techniques et de l'interaction entre science et technologie[29]. Nous pouvons donc considérer que la scientométrie comme un champ de la bibliométrie spécialisée qui s'intéresse au domaine de l'information scientifique et technique (IST).

Selon Jean-Pierre Courtial, la bibliométrie est une approche quantitative des techniques de gestion d'une bibliothèque et au comptage de tout ce que l'on peut trouver dans une bibliothèque scientifique. Alors que la scientométrie pour lui renvoie à la généralisation de ces techniques qui incluent non seulement les documents publiés, mais

26 Michel Callon et al., *La scientométrie*, 1993, p. 3.

27 Jean-Max Noyer, *Les sciences de l'information – Bibliométrie, Scientométrie, Infométrie*, 1995, p. 14.

28 Jean-Max Noyer, *Les sciences de l'information – Bibliométrie, Scientométrie*, Infométrie, 1995, p. 16.

29 Jean-Max Noyer, *Les sciences de l'information – Bibliométrie, Scientométrie, Infométrie*, 1995, p. 15.

aussi les citations qu'ils ont reçu afin de gérer l'activité de recherche scientifique[30]. Yves Gingras fait également la distinction entre ces deux concepts. Selon Gingras, la scientométrie porte sur la mesure quantitative de l'ensemble des activités scientifiques, toutes disciplines confondues. Il souligne aussi que la scientométrie s'appuie sur des données portant à la fois sur les investissements pour la recherche et le développement, la formation du personnel scientifique et la production d'articles et de brevets. Tandis que la bibliométrie est un sous-ensemble de la scientométrie qui se limite uniquement à l'analyse des publications et de leurs propriétés[31].

Enfin, le terme « infométrie » a été employé initialement en 1979 par un documentaliste allemand, Otto Nacke. Il a été ensuite adopté par la « Fédération internationale de documentation » (FID) pour désigner l'ensemble des activités métriques relatives à l'information, couvrant aussi bien la bibliométrie que la scientométrie[32]. Selon Fidelia Ibekwe-SanJuan, l'infométrie renvoie à « *l'application des méthodes bibliométriques au traitement de l'information de tout type, non seulement bibliographique, mais également au contenu des publications et de la communication de l'information.*[33] » Pour elle, l'infométrie incluent à la fois la bibliométrie et la scientométrie. Nous pouvons donc considérer que l'infométrie comme un domaine très large qui ne se réduit pas seulement aux données scientifiques et qu'il comprend d'autres disciplines telles que la webométrie (= c'est l'application des méthodes bibliométriques à l'univers du web). Jean-Max Noyer, définit l'infométrie comme un « *ensemble d'activités dans le domaine de la documentation et en particulier de l'information scientifique et techniques.*[34] »

Pour résumer, nous considérons que ces trois notions : bibliométrie, scientométrie et infométrie reposent sur les mêmes méthodes. La bibliométrie et la scientométrie visent avant tout à mesurer la science. La première s'intéresse principalement à la mesure de la production scientifique. Et la seconde s'applique aussi bien au domaine scientifique et technique qu'aux ressources humaines, économiques, etc. L'infométrie couvre, quant à elle, un champ disciplinaire très vaste et englobe la bibliométrie et scientométrie. Enfin, nous pouvons constater que plusieurs approches montrent un passage de la bibliométrie à l'infométrie en passant par la scientométrie.

30 Jean-Pierre Courtial, *Introduction à la scientométrie : de la bibliométrie à la veille technologique*, 1990, p. 7.

31 Yves Gingras, *Les dérives de l'évaluation de la recherche – Du bon usage de la bibliométrie*, 2013, p. 9.

32 Jean-Max Noyer, *Les sciences de l'information – Bibliométrie, Scientométrie, Infométrie*, 1995, p. 16.

33 Fidelia Ibekwe-SanJuan, *La science de l'information : origines, théories et paradigmes*, 2012, p. 182.

34 Jean-Max Noyer, *Les sciences de l'information – Bibliométrie, Scientométrie, Infométrie*, 1995, p. 15.

3 Les laboratoires de recherche en SIC :

Dans cette dernière partie de ce premier chapitre, nous allons tenter de définir ce qu'un laboratoire de recherche en Sciences de l'information et de la communication (SIC). Pour ce faire, nous nous focaliserons sur la question des frontières d'un laboratoire de recherche en SIC. Mais avant d'aborder ce sujet, il nous semble qu'il est important de rappeler, sans rentrer dans les détails de son histoire, la genèse et une ou deux définitions des Sciences de l'information et de la communication.

Nous pouvons tout d'abord souligner que les Sciences de l'information et de la communication sont une spécificité française dès le début des années 1970. Car dans les autres pays, il y a une séparation entre les deux branches qui constituent la discipline alors qu'en France, nous avons réuni les SIC dans une seule discipline que l'on considère souvent comme une inter-discipline.

En effet, Les SIC sont le fruit de la fusion qui a eu lieu dans les années 1970 entre la « Science de l'information » et la « Science de la communication ». Cette association des deux disciplines était une initiative d'un groupe de trois chercheurs de différents champs disciplinaires : Jean Meyriat, Roland Barthes et Robert Escarpit. Ils ont donc réuni d'autres chercheurs pour constituer un Comité des Sciences de l'information et de la communication[35]. Lors de la création de ce Comité des SIC en 1972, Jean Meyriar a expliqué que « *le terme des SIC est finalement conservé pour des raisons d'efficacité : le sentiment prévaut que le mot plus concret d'information précise un peu la notion vague de communication ; ce couplage permet en même temps de servir les intérêts de plusieurs groupes distincts de spécialistes, sans prendre une position définitive sur l'épistémologie du domaine.*[36] » Cela montre l'importance et l'intérêt de fusionner entre deux branches pour en faire une seule discipline. Les SIC sont une inter-discipline se caractérisant par une dualité entre information et communication. Leur objet scientifique n'est pas clairement défini.

Après la reconnaissance institutionnelle des SIC et la création de la 71ème section par le Conseil national des universités (CNU), les Sciences de l'information et de la communication sont définis comme une « *inter-discipline centrée sur l'étude des processus de l'information et de la communication relevant d'actions organisées, finalisées, prenant ou non appui sur des techniques, et participant des médiations sociales et culturelles* » (CNU 71ème section, 1993)[37]. Olivesi définit également les SIC comme une inter-discipline. Il écrit notamment que « *(...)les SIC sont se positionnent comme une inter-discipline capable de traiter de problématiques que les autres disciplines propres aux SHS délaissent ou ne traitent que partiellement.*[38] » Nous pouvons donc le constater à

35 Remplacé par « la Société Française des Sciences de l'Information et de la Communication » (SFSIC)

36 Citation de l'ouvrage dirigé par R. Boure, « *Les origines des Sciences de l'information et de la communication : Regards croisés* », p.10.

37 Hubert Fondin, « *La science de l'information ou le poids de l'histoire* », 2006, p. 1.

38 Olivesi, « Sciences de l'information et de la communication », 2006, p. 194.

travers ces définitions que les Sciences de l'information et de la communication sont interdisciplinaires.

C'est cette question d'interdisciplinarité des SIC qu'il faut penser pour définir les frontières de la discipline et de ses laboratoires de recherche. S'interroger sur les frontières d'un laboratoire en SIC, c'est d'abord questionner sur les liens que les SIC nouent avec les autres disciplines. Elles s'appuient généralement sur des cadres théoriques et méthodologiques issus de divers disciplines telles que la sociologie, la linguistique, la philosophie, etc. Cependant, nous pouvons constater que la question des frontières disciplinaires est assez complexe. Dès lors que l'on s'interroge sur les frontières disciplinaires des SIC, elles nous paraissent compliquées à délimiter, et plus on les approche pour tenter de les définir plus elles deviennent floues.

Selon Mélanie Bourdaa et Aurélia Lamy, les laboratoires de recherche en Sciences de l'information ont un rôle important dans le progrès des SIC tant au niveau national qu'international. Elles les définissent comme des acteurs majeurs du développement des Sciences de l'information et de la communication en France et à l'étranger. Ce sont des lieux d'accueil privilégiés des chercheurs en SIC, des espaces de formation pour les doctorants, des espaces de diffusion du savoir et des lieux d'échange institutionnel[39]. Les laboratoires de recherche en SIC s'ouvrent de plus en plus à d'autres disciplines. Nous constatons que pour des raisons généralement économiques, certains laboratoires en SIC tentent davantage de se regrouper avec d'autres laboratoires ou équipes de recherches. Ces regroupements présentent à la fois un avantage et aussi un inconvénient pour les équipes ou les laboratoires en SIC. L'avantage de ces processus de fusions disciplinaires entre les équipes de recherche est qu'elles permettent d'élargir le périmètre et les champs de recherche des SIC, d'ouvrir la discipline des SIC à de nouvelles problématiques et de nouvelles méthodologies ainsi que de renforcer leur interdisciplinarité. Cependant, l'inconvénient est que cela conduit généralement les équipes et les laboratoires de recherche à redéfinir leurs frontières disciplinaires ainsi qu'à se poser des questions sur leur appartenance et leur part de production scientifique.

En effet, nous constatons qu'après la fusion de certains laboratoires en SIC avec d'autres entités de recherche, les laboratoires de recherche en SIC sont envahis par des chercheurs d'autres branches. Comme par exemple le CREM (Centre de recherche sur les médiations) qui s'est associé avec des chercheurs de Sciences du langage et d'autres disciplines. Inversement, nous avons des équipes de recherche en SIC qui intègrent à des laboratoires appartenant à d'autres disciplines. Nous pouvons citer ici le CRAPE (Centre de recherches sur l'action politique en Europe)), un laboratoire de science politique qui est devenu une Unité mixte de recherche (UMR) en sciences humaines et sociales (ARENES) après avoir intégré des équipes de recherche relevant d'autres disciplines telles que les Sciences de l'information et de la communication.

39 Mélanie Bourdaa et Aurélia Lamy, « *Les laboratoires de recherche en Sciences de l'information et de la communication* », 2013, p.1.

Ce que nous voulions par montrer par ces exemples, c'est que la fusion des équipes ou des laboratoires de recherche ainsi que la création des UMR pose un énorme problème dans la définition des frontières d'un laboratoire en SIC. Nous pouvons en effet souligner le problème d'appartenance disciplinaire même si, a priori, cela ne pose aucun problème pour certaines unités de recherche. Ces associations et ces regroupements des équipes et des unités de recherche en SIC rendent difficile de définir réellement ce qu'un laboratoire de recherche en Sciences de l'information et de la communication, car leurs frontières sont en pleine mutation.

Les Sciences de l'information et de la communication sont évidemment interdisciplinaires et c'est cette interdisciplinarité qui les constitue, mais on se pose beaucoup de questions sur cette notion. Car il est vrai que l'interdisciplinarité est fondamentale pour notre domaine des SIC, mais on nous reproche souvent de ne pas avoir spécifié ce qui nous distingue des autres disciplines voisines. Selon Mélanie Bourdaa et Aurélia Lamy, l'interdisciplinarité est à la fois une force et une faiblesse pour notre discipline. Car pour elles et selon le point de vue de l'épistémologie, « *il est intéressant et enrichissant de puiser dans plusieurs camps disciplinaires. Cependant, il est souvent reproché aux SIC de ne pas avoir d'ancrage scientifique propre à la discipline.*[40] »

L'évolution des frontières disciplinaires des laboratoires de recherche en SIC nous conduit à souligner la richesse et la diversité des thématiques, des méthodologies ainsi que de la production des laboratoires de notre discipline. Les laboratoires en SIC s'ouvrent de plus en plus à d'autres disciplines. Cela ne fait que renforcer l'interdisciplinarité. Cependant, nous pouvons relever de nombreuses questions sur cette ouverture et ces mutations des frontières de nos laboratoires de recherche. Nous nous interrogeons notamment sur le degré de représentativité de la discipline par certains laboratoires reconnus comme des laboratoires de recherche relevant des Sciences de l'information et de la communication. L'intégration des équipes de recherche d'autres disciplines au sein d'un laboratoire en SIC pourrait également fragiliser la position du laboratoire dans la discipline et remettre en question son inscription dans le domaine des Sciences de l'information et de la communication.

40 Mélanie Bourdaa et Aurélia Lamy, « *Les laboratoires de recherche en Sciences de l'information et de la communication* », 2013, p.3.

II. Cadre méthodologique : objectifs, méthodologie et questions méthodologiques

Dans la section précédente, nous avons tenté de faire un tour d'horizon des études similaires à la nôtre en présentant les auteurs et rappelant le contexte de leurs études, les objectifs, certains résultats et conclusions des auteurs. Nous avons également rappelé les différentes théories et approches ayant travaillé et tenté de définir les concepts de bibliométrie, scientométrie et infométrie. Ainsi, nous avons évoqué la notion des frontières disciplinaires d'un laboratoire en Sciences de l'information et de la communication. Intéressons-nous maintenant au cadre méthodologique de notre étude. Dans cette section, nous allons dans un premier temps évoquer les principaux objectifs de notre analyse. Dans un second temps, nous détaillerons la méthodologie mise en place pour cette étude. Enfin, dans un troisième temps, nous aborderons les différents problèmes méthodologiques rencontrés lors de la constitution du corpus.

1 Objectifs de l'étude :

Cette étude est centrée sur l'analyse des publications présentes dans les collections des laboratoires de recherche en Sciences de l'information et de la communication sur la plateforme HAL. L'objectif est d'étudier la présence des laboratoires de recherche de notre domaine des SIC sur HAL en fonction de nombreux indicateurs et variables : nombre de dépôts et part des documents en libre accès, nombre de documents par genres et la langue de publication.

Notre approche s'inscrit dans une perspective bibliométrique. En effet, à partir des données quantitatives que nous avons recueilli sur HAL via l'API de recherche de HAL, nous analyserons l'évolution des dépôts sur HAL des laboratoires sélectionnées. Nous étudierons aussi le type de dépôts et la part des documents en libre accès. Nous nous intéressons également à la langue des documents afin de déterminer la part des publications en langue étrangère dans le but est de mesurer l'internationalisation des collections. Ainsi, nous tenterons d'analyser la typologie des documents que les laboratoires de recherche en SIC publient dans leurs collections.

2 Méthodologie :

La méthodologie mise en place pour cette étude repose sur trois grandes étapes importantes pour constituer un corpus de données propre. En effet, il s'agit dans un premier temps de constituer une liste raisonnée des laboratoires de recherche en Sciences de l'information et de la communication. Dans un second temps, extraire de façon automatique des données à partir des API de HAL. Et dans un dernier temps, il nous a fallu vérifier, nettoyer et traiter manuellement les données collectées pour pouvoir enfin avoir un corpus convenable.

2.1 Constitution du panel : liste des laboratoires en SIC

La première tâche a consisté à recenser l'ensemble des laboratoires de recherche en Sciences de l'information et de la communication, que ce soit un laboratoire purement ou partiellement en SIC. Le principe qui a présidé au recensement des laboratoires de recherche était d'établir une liste raisonnée, c'est-à-dire une liste bien réfléchie au préalable comprenant le plus possible d'informations pertinentes sur chaque laboratoire.

Pour élaborer une liste exhaustive de 36 laboratoires de recherche en SIC, nous nous sommes appuyés sur trois listes différentes préétablies comprenant 70 laboratoires au total. La première liste comporte 30 laboratoires, membres de la CPDirSIC[41] que nous avons donc récupérée sur le site de la CPDirSIC[42]. La deuxième liste a été établie par le Ministère de l'enseignement supérieur[43]. Cette liste comprenait 38 laboratoires, mais 6 laboratoires ont été déjà identifiés dans la liste de la CPDirSIC. La dernière liste a été établie, lors d'une réunion des membres de la CPDirSIC, par un chercheur en SIC dont nous ignorons le nom. Cette dernière liste comptait 39 laboratoires, 31 laboratoires ont été déjà identifiés dans les deux premières listes.

À partir de cette liste de 70 laboratoires de recherche, nous avons d'abord établi une première liste de 41 laboratoires. Nous avons en effet constaté que la liste du Ministère de l'enseignement supérieur confondait entre les laboratoires des Sciences de l'information et de la communication (SIC) et les laboratoires des Sciences et technologies de l'information et de la communication (STIC). Il y avait également certains laboratoires qui se sont regroupés sous un même nom après la fusion de leurs unités de recherche. Puis, nous avons établi une liste finale comprenant 36 laboratoires de recherche représentatifs des SIC. Nous avons donc éliminé 5 laboratoires de recherche que l'on considère comme « non représentatifs des SIC », car le nombre de chercheurs en SIC

41 Conférence Permanente des Directeurs.trices de laboratoires en Sciences de l'Information et de la Communication

42 CPDirSIC. « *Membres de la CPDirSIC* ». Consulté 5 décembre 2020. http://cpdirsic.fr/membres-de-la-cpdirsic/

43 Ministère français de l'Enseignement supérieur, de la Recherche et de l'Innovation « *Liste des Unités de Recherche, de la discipline : Sciences de l'information et de la communication* ». Consulté 5 décembre 2020. https://appliweb.dgri.education.fr/annuaire/ListeEntite.jsp?entite=ur&sd=22&prov=MotCle

des cinq laboratoires exclus était très faible et ne comptaient dans leurs équipes que deux ou trois membres en Sciences de l'information et de la communication.

En ce qui concerne nos principales sources d'informations pour la collecte des informations sur les laboratoires de recherche, nous nous sommes servis principalement du répertoire national des structures de recherche[44] (RNSR) qui contient de nombreuses informations sur les structures de recherche. Nous nous sommes aussi appuyés sur le moteur de recherche scanR[45] pour vérifier certaines informations ou d'en trouver éventuellement d'autres. Nous avons également utilisé les sites des laboratoires, le site du Hcéres[46] (Haut Conseil de l'évaluation de la recherche et de l'enseignement supérieur) ainsi que la plateforme AURéHAL[47] (Accès Unifié aux Référentiels HAL) qui nous a permis de récupérer l'identifiant HAL de structure via le référentiel des structures de recherche.

Pour chaque laboratoire, nous avons dans un premier temps déterminé l'adresse du site web, l'acronyme, le nom, la discipline (laboratoire purement SIC, interdisciplinaire, pluridisciplinaire), le type d'unité (UMR, ULR, UPR,…), le numéro d'unité, l'année de création, le numéro identifiant national RNSR, le domaine scientifique RNSR et le classement scientifique ERC. Dans un second temps, nous avons déterminé les axes et les thématiques de recherche, la ou les tutelle(s), le responsable de l'unité et les coordonnées de contact (numéro de téléphone, adresse électronique ainsi que postale). Dans un dernier temps, nous avons également identifié pour chaque laboratoire l'identifiant de structure HAL et l'adresse de la collection HAL de côté usager.

2.2 Collecte et extraction des données à partir de HAL :

Une fois que nous avons constitué la liste des laboratoires de recherche et déterminé toutes les informations que l'on souhaitait obtenir, nous avons procédé à l'extraction des données à partir de HAL en utilisant l'interface de programmation d'applications (API).

Avant de décrire comment nous avons procédé à l'extraction des données à partir de HAL, les requêtes utilisées et le corpus de données que nous avons pu obtenir grâce aux API-HAL. Il nous semble qu'il est d'abord intéressant de définir brièvement ce qu'est une API et de présenter l'API de recherche de la plateforme HAL.

2.2.1 Présentation de l'API de recherche de HAL :

Une API désigne « *Application Programming Interface* » que l'on traduit en français par interface de programmation d'applications. Il s'agit d'un système qui relie divers programmes entre eux et leur permet de communiquer ensemble, de transmettre et d'échanger des données. Une API sert de point d'entrée à un autre logiciel. Selon Yves

44 RNSR : https://appliweb.dgri.education.fr/rnsr/ChoixCriteres.jsp?PUBLIC=OK

45 ScanR : https://scanr.enseignementsup-recherche.gouv.fr/

46 Hcéres : https://www.hceres.fr/fr/rechercher-une-publication?key=&f%5B0%5D=themes_publications%3A43

47 AURéHAL : https://aurehal.archives-ouvertes.fr/structure

Tomic, « *une API permet à un système source d'exposer ses données à des fins d'exploitation par d'autres systèmes*[48] ».

L'API de recherche de HAL est un des services offerts par cette plateforme d'archive ouverte. Il s'agit d'une façade permettant une interaction machine à machine. Elle offre un accès à une multitude de données et d'informations disponibles sur HAL. En effet, grâce à cette interface, nous pouvons accéder et extraire de la façon automatique l'ensemble des métadonnées des publications déposées dans la plateforme HAL. Elle permet également de nombreuses fonctionnalités telles que l'utilisation des facettes, le tri des résultats, le choix d'un format de réponse, les filtres sur les requêtes, l'utilisation des opérateurs booléens, spécifier les champs à retourner dans la réponse, etc.

Tout d'abord, le point d'accès de l'API de recherche de HAL s'effectue dans le navigateur via l'URL suivante : < https://api.archives-ouvertes.fr/search/ >. À partir de cette URL, nous pouvons limiter notre requête soit à un portail (ex : archivesic, TEL,...) ou à une collection donnée. Il suffit seulement le préciser dans l'URL comme dans les exemples suivants : < https://api.archives-ouvertes.fr/search/archivesic/ > si on veut limiter la requête au portail « *archivesic* » ou bien < https://api.archives-ouvertes.fr/search/GERIICO/ > si nous souhaitons limiter notre requête à la collection du laboratoire GERiiCO.

Ensuite, concernant la syntaxe d'une requête, seul paramètre obligatoire dans une requête est le paramètre « q » que l'on met toujours au début de la requête. Ce paramètre doit être précédé d'un point d'interrogation « ? ». Chaque paramètre que l'on intègre dans une requête doit avoir une valeur (paramètre = valeur). Nous pouvons attribuer à ce premier paramètre obligatoire une valeur par défaut qui est « *.* » (ex :?q=*.*). Ainsi, nous pouvons insérer plusieurs paramètres dans une même requête, mais nous devons les séparer par le caractère « & », comme nous pouvons le voir dans l'exemple suivant : < https://api.archives-ouvertes.fr/search/GERIICO/?q=*.*&wt=csv&rows=100 >.

Pour terminer, nous pouvons citer quelques fonctionnalités de l'API de recherche de HAL. Parmi ces fonctionnalités, nous retrouvons la fonctionnalité permettant de choisir un format de sortie. En effet, le format de sortie/réponse par défaut est « json », mais plusieurs formats sont mis à disposition par l'API de recherche de HAL. En utilisant le paramètre « wt », nous pouvons choisir un autre format de réponse parmi les huit formats proposés (csv, xml, xml-tei,…). Une autre fonctionnalité intéressante est celle qui permet de spécifier les champs à retourner dans la réponse grâce au paramètre « fl ». Ce paramètre est très avantageux, car il peut prendre plusieurs valeurs en même temps, mais elles doivent être séparées par une virgule « , ». La seule chose que l'on pourrait reprocher à cette fonctionnalité est qu'il n'est disponible qu'avec trois formats de sortie : json, xml et csv. L'API-HAL offre également la possibilité d'effectuer de nombreux filtres sur les requêtes en utilisant le paramètre « fq » (Exemple pour retourner uniquement les dépôts avec fichier : fq=submitType_s:file). Enfin, il existe de nombreux paramètres et

48 Tomic Yves, « *De l'usage des API. Les API de l'Abes* », 2014. p. 17. https://www.cairn.info/revue-documentaliste-sciences-de-l-information-2014-3-page-17.htm

fonctionnalités que nous ne pouvons pas tous citer ici. Cependant, nous pouvons souligner que les équipes de la plateforme HAL ont mis à disposition toute une documentation[49] que l'on trouve très bien détaillée pour maîtriser cette API.

2.2.2 Requêtes et interrogation de l'API :

Après avoir pris connaissance de la documentation de l'API de recherche de HAL et maîtrisé la syntaxe des requêtes de recherche, nous avons dans un premier temps identifié pour chaque laboratoire de recherche l'adresse de la collection HAL de côté de l'API. En effet, comme nous l'avons mentionné dans la partie précédente que le point d'entrée de cette API se fait à partir du navigateur, c'est pour cela qu'il nous ait donc fallu de déterminer l'URL-API pour chaque collection. Nous avons principalement utilisé le code de collection (l'acronyme du laboratoire) pour les identifier, sauf pour une minorité des laboratoires qui ne possèdent pas une collection. Nous avons utilisé le code de structure HAL pour ces quelques laboratoires qui n'ont pas une collection sur HAL.

Une fois que l'identification des collections sur l'interface de recherche de HAL est achevée, nous avons procédé à la construction des requêtes afin d'interroger les collections des laboratoires sélectionnées et d'extraire les données dont nous avons besoin pour notre analyse.

Notre méthode de collecte des données s'appuie sur un ensemble de requêtes permettant d'interroger, via l'API, le serveur de la plateforme HAL et d'en extraire les données demandées dans chaque requête. L'extraction des données porte sur la période 2001-2021.

Dans un premier temps, nous avons élaboré une première requête permettant de filtrer les données sur un intervalle d'années. Ce qui nous a permis de faire des extractions année par année (de 2001 à 2021) à partir de chaque collection. Grâce à cette requête, nous avons pu obtenir, en format CSV, l'ensemble des données concernant le nombre de documents par année et par genres (article, communication, thèse, etc.).

Dans un deuxième temps, nous avons utilisé une autre requête pour pouvoir extraire le type de dépôt (fichier, notice, annexe) afin d'avoir le nombre de notices, d'annexes et des documents disponibles avec fichier, c'est-à-dire le nombre de dépôts en texte intégral qui nous servira comme indicateur de la part des documents en libre accès.

Ensuite, nous nous sommes servis d'une autre requête avec laquelle nous avons pu récupérer les données concernant la langue de publication de chaque document.

Nous avons également construit une dernière requête dans laquelle nous avons précisé, grâce au paramètre « fl », un certain nombre important de champs à retourner dans un seul fichier CSV que l'on a enregistré sur notre machine. Cela nous a permis d'avoir un seul fichier par collection avec toutes les données dont nous avons besoin pour

49 CCSD « *API HAL | API de recherche HAL* ». Consulté 2 juin 2021. https://api.archives-ouvertes.fr/docs/search

cette première partie de notre étude ainsi que de sauvegarder les autres données qui nous serviront plus tard pour la deuxième partie de l'étude, qui portera sur l'analyse des thématiques des publications, l'origine des contributeurs et les réseaux de collaboration. Ce fichier comporte les champs suivants : *version du document, type de documents, titre, nom(s) de d'auteur(s), acronyme laboratoire(s), acronyme institution(s), citation abrégée, titre de revue, type de dépôts, année de dépôt, mots-clés, langue de document, nom du contributeur/déposant, regroupement de laboratoire (acronyme), regroupement d'institutions (acronyme)*.

Grâce à ces extractions automatiques des données via l'API-HAL et la possibilité d'exporter les données en format CSV, l'API de recherche de HAL nous a beaucoup facilité cette tâche. Cependant, toutes les données recueillies via l'API de HAL ont subi, à l'aide des fonctionnalités de filtre et de tri sur Excel, une longue opération de nettoyage, du traitement, du comptage et du travail manuel afin d'obtenir un corpus propre qui nous permettra d'effectuer notre analyse.

2.2.3 Présentation du corpus des données :

Comme mentionné plus haut, notre étude est basée sur les publications effectuées dans les collections HAL des laboratoires en Sciences de l'information et de la communication. Elle a été menée à partir d'un échantillon représentatif des laboratoires de recherche en SIC et repose sur des données extraites de la plateforme HAL. En effet, le corpus est composé de 36 laboratoires de recherche en SIC et de 35 381 dépôts recueillis à partir des collections HAL de ces laboratoires en utilisant des extractions via l'API de recherche de HAL. Notre corpus comprend donc des données quantitatives des dépôts effectuées dans les collections HAL des laboratoires en SIC pour la période 2001-2021. Les données sont au format CSV. Ce qui nous a permis et facilité de nettoyer ainsi que de filtrer les données selon le type de documents (articles, communications, chapitres d'ouvrages, thèses, etc.), le type de dépôts (notices, texte intégral et annexes), la langue de communication (français, anglais et autres) ainsi que l'année de publication.

Tableau 1: Récapitulatif de la présentation du corpus

Périmètre	Extraction de données des collections HAL des laboratoires en SIC
Provenance	Extraction via l'API-HAL
Format	CSV
Période	2001-2021
Volume	35 381 dépôts
Types de documents	19 types de documents
Type de dépôts	Notices, texte intégral et annexes
Langue	Français, anglais et autres
Autres métadonnées	Titre, auteurs, mots-clés, années de dépôt, version du document, contributeur, acronyme laboratoire et institution, etc.

2.3 Questions méthodologiques : difficultés rencontrées

Lors de la constitution de notre corpus, nous avons été confrontés à quelques difficultés et questions méthodologiques pour établir un corpus propre et complet.

En effet, durant la première étape concernant l'élaboration d'une liste des laboratoires de recherche en SIC, nous avons constaté que les données de certains laboratoires ne sont pas toujours à jour. Il nous aurait fallu d'être constamment attentif à la fiabilité des données que nous trouvons. Nous avons dû vérifier sur les différents sites utilisés avant de collecter les données dont nous pensons qu'elles ne sont pas correctes. Nous avons également observé qu'il existe plusieurs codes de structure pour un même laboratoire de recherche. Nous les avons tous collectés et testés lors des extractions de données. Nous nous sommes aperçus que certains de ces codes de structure ne sont plus valables ou bien contiennent seulement très peu de publications, c'est pourquoi nous avons privilégié d'utiliser le code de collection (acronyme du laboratoire) pour les extractions des données.

Un autre point important à souligner, il s'agit de la question des frontières d'un laboratoire de recherche en SIC. L'interdisciplinarité des laboratoires de recherche constitue un frein et nous a rendu la tâche plus difficile, car nous avons constaté que certains laboratoires de recherche (UMR) sont plus compliqués que d'autres. Car la fusion de certaines équipes de recherche en SIC avec d'autres unités relevant d'autres disciplines nous conduit à nous poser la question sur leur appartenance disciplinaire et nous demander le laboratoire s'inscrit véritablement dans le champ des SIC. Nous pouvons donc dire que l'appartenance disciplinaire de certains laboratoires de recherche en SIC est floue et que leur l'interdisciplinarité rend difficile de définir et de délimiter leurs frontières.

Enfin, malgré la richesse des API de HAL, nous pouvons souligner quelques problèmes. En effet, pendant les extractions des données à partir de l'API, nous avons été confrontés à un problème lié à l'incomplétude des données. Nous avons constaté que parfois le champ des mots-clés est retourné vierge pour certains documents. De même pour le champ contenant l'affiliation des auteurs, il est aussi incomplet et manque des données par rapport à ce que l'on voit sur l'interface usager. Cela demande donc beaucoup de temps et un travail manuel colossal. Comme nous l'avons déjà mentionné, il existe plusieurs codes de structure pour un seul laboratoire. Nous avons également relevé qu'il y a une différence entre la nomenclature HAL et celle du HCERES. Cela pose un vrai problème pour les laboratoires et les chercheurs qui publient sur HAL, car ils sont obligés d'obéir à la nomenclature imposée par le HCERES dans leurs bilans. La question d'homogénéisation de la nomenclature HAL et HCERES est importante, car cela pourrait inciter les chercheurs à déposer leurs productions dans HAL. Cette différence de nomenclature ne constitue pas vraiment un véritable problème pour notre étude ni pour HAL, mais il s'agit plutôt d'une question générale que nous avons voulu signaler. C'est d'ailleurs une des questions que l'on étudiera plus tard, dans la continuité de ce projet.

III. Présentation et analyse des résultats

Dans cette dernière section, nous allons exposer les résultats de notre étude. Dans un premier temps, nous nous intéressons à la question de présence des laboratoires de recherche en SIC sur HAL dont nous aborderons le nombre de laboratoires qui disposent d'une collection HAL, l'année du début de dépôts sur HAL et le nombre de dépôts par laboratoires. Dans un second temps, nous traiterons la question de la typologie des documents déposés dans les collections des laboratoires. Dans un troisième temps, nous analyserons la part des documents en libre accès. Ensuite, nous tenterons de déterminer le degré d'ouverture internationale des collections à travers la langue de publication. Enfin, juste avant la conclusion, nous terminerons cette section par une discussion des résultats.

1 Présence des laboratoires de recherche en SIC sur HAL :

L'ensemble des laboratoires de recherche en SIC que nous avons sélectionné sont présents sur HAL. En effet, ils possèdent tous un identifiant HAL (code de structure) dans le référentiel de recherche[50]. Cependant, certains laboratoires de recherche n'ont pas une collection HAL, mais disposent seulement d'un identifiant HAL. Cela ne représente qu'une petite minorité des laboratoires de recherche choisis. Seulement 6 laboratoires qui ne possèdent pas une collection contre 30 laboratoires de recherche qui ont une collection sur HAL (Figure 1).

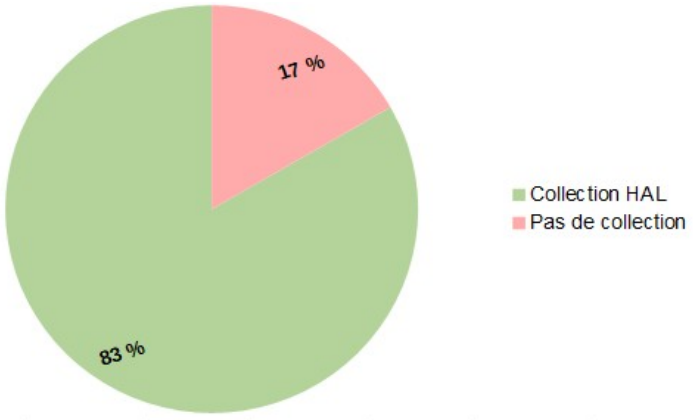

Figure 1: Collection HAL des laboratoires de recherche en SIC (N=36 labos)

50 AURéHAL : https://aurehal.archives-ouvertes.fr/structure

Concernant le début des dépôts, tous les laboratoires n'ont pas commencé à la même période (Figure 2). Nous pouvons distinguer trois périodes essentielles ayant marqué le début des dépôts sur HAL par les laboratoires. La première période remonte aux cinq premières années après la création de HAL (2001-2005) dont nous avons 16 laboratoires qui ont commencé à faire leurs premiers dépôts sur HAL. La deuxième est celle de 2006 à 2010 dont nous avons compté 14 nouveaux laboratoires qui ont aussi débuté à déposer sur HAL. La dernière s'étend de 2011 à 2021 avec 6 nouveaux laboratoires. Ces chiffres pourraient bien expliquer l'intérêt des laboratoires de recherche en SIC pour la plateforme HAL. Les 6 laboratoires qui ont débuté lors de la deuxième décennie de la mise en place de HAL sont généralement des nouveaux laboratoires créés après 2010, ou bien, ils se sont fusionnés avec d'autres unités de recherche et auraient donc dû, peut-être, changer leur code de structure et perdre leurs dépôts antérieurs.

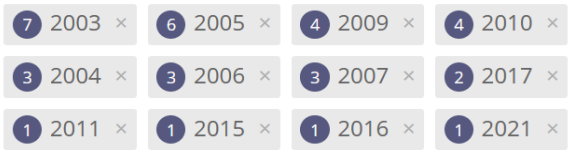

Figure 2: Nombre de laboratoires par année du début de dépôts (N=36 labos)

Les 36 laboratoires de recherche totalisent, en mai 2021, 35 381 dépôts sur leurs collections HAL (79 % des dépôts en SIC si l'on compare au nombre de dépôts dans le portail « ArchiveSIC[51] »). La moyenne des dépôts par laboratoires est de 983 dépôts. Mais cette moyenne n'est pas vraiment juste et induit à l'erreur, car le nombre de dépôts pour chaque laboratoire est très variable, comme nous pouvons le voir dans la « figure 3 » qui montre une répartition très irrégulière : très forte, moyenne, faible et très faible.

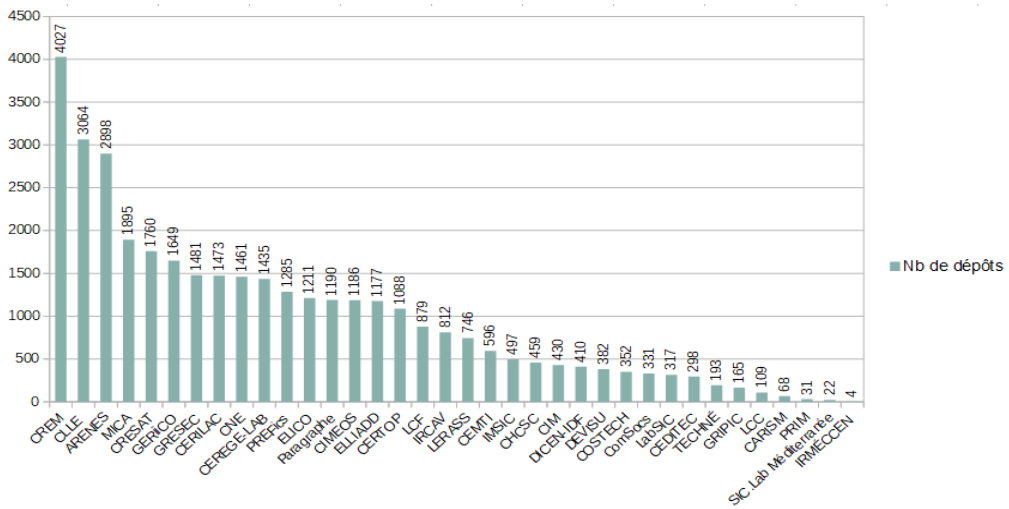

Figure 3: Nombre de dépôts sur HAL par laboratoire (N=35 381)

51 44964 dépôts, en mai 2021, dans « ArchiveSIC » https://hal.archives-ouvertes.fr/search/index/?q=%2A&domain_t=shs.info

Nous pouvons constater dans la « figure 3 » que certains laboratoires produisent beaucoup de publications sur HAL ; des laboratoires produisant un peu moins et qui rentrent dans la moyenne ; d'autres en dessous de la moyenne ; alors qu'il y a des laboratoires qui ne publient quasiment pas sur HAL. Nous pouvons voir dans cette distribution que le premier laboratoire est le « CREM » (*Centre de recherche sur les médiations*) avec plus de 4 000 dépôts, suivi par « CLLE » (*Laboratoire Cognition, Langues, Langage, Ergonomie*) avec plus de 3 000 dépôts, puis ARENES (*Centre de recherches sur l'action politique en Europe*) qui totalise un peu moins de 3 000 dépôts. Ces trois laboratoires sont des unités mixtes de recherche (UMR) qui réunissent un nombre important de chercheurs et d'équipes de différentes disciplines. Ce qui pourrait peut-être expliquer ce nombre important des dépôts dans leurs collections. Nous pouvons donc distinguer quatre catégories de laboratoires (Tableau 2).

Tableau 2: Répartition des laboratoires selon leur nombre de dépôts

	Moyenne des dépôts
16 laboratoires ont **moins de 499 dépôts**	**254**
04 laboratoires ont **entre 500 et 999 dépôts**	**758**
13 laboratoires ont **entre 1 000 et 1 999 dépôts**	**1 407**
03 laboratoires ont **plus de 2 000 dépôts**	**3 330**

Nous pouvons interpréter cette distribution de deux façons différentes : en nous appuyant soit sur ce que l'on appelle « longue traîne » en statistique, soit sur la « loi de Pareto ». Selon la longue traîne, nous pouvons constater que 8 % des laboratoires (3) ont produit 28 % des dépôts sur HAL, 36 % des laboratoires (13) ont produit 52 % des dépôts et 56 % des laboratoires (20) ont produit seulement 20 % des dépôts. Alors selon le principe de Pareto, nous pouvons remarquer que 44 % des laboratoires de recherche (16) ont produit 80 % des dépôts sur HAL, tandis que 56 % des laboratoires (20) ont produit 20 % des dépôts. Ce qui se traduit parfaitement par la loi des 80-20.

En ce qui concerne l'évolution de la somme des dépôts déposés par l'ensemble des laboratoires de recherche en SIC dans HAL (Figure 4), nous pouvons observer une évolution faible entre les années 2003 et 2006. Le nombre de dépôts pour cette période ne dépasse pas 300 dépôts. Puis, une évolution assez importante de 2006 à 2007 passant de 289 à 807 dépôts. Cette évolution va continuer de manière constante jusqu'à 2009 pour arriver aux 1 174 dépôts avec une moyenne de 180 dépôts par année. Nous pouvons ensuite constater une nouvelle évolution considérable de 2009 à 2010. Le nombre de dépôts a quasiment doublé par rapport à l'année précédente. Le nombre des dépôts va poursuivre son évolution de façon constante aussi jusqu'à 2014 avec une moyenne de 1 330 à 1 500 dépôts par année. Enfin, une dernière évolution marquante de 2014 au mi-semestre de l'année 2021 dont nous notons une évolution moyenne à environ 4 000 dépôts par année.

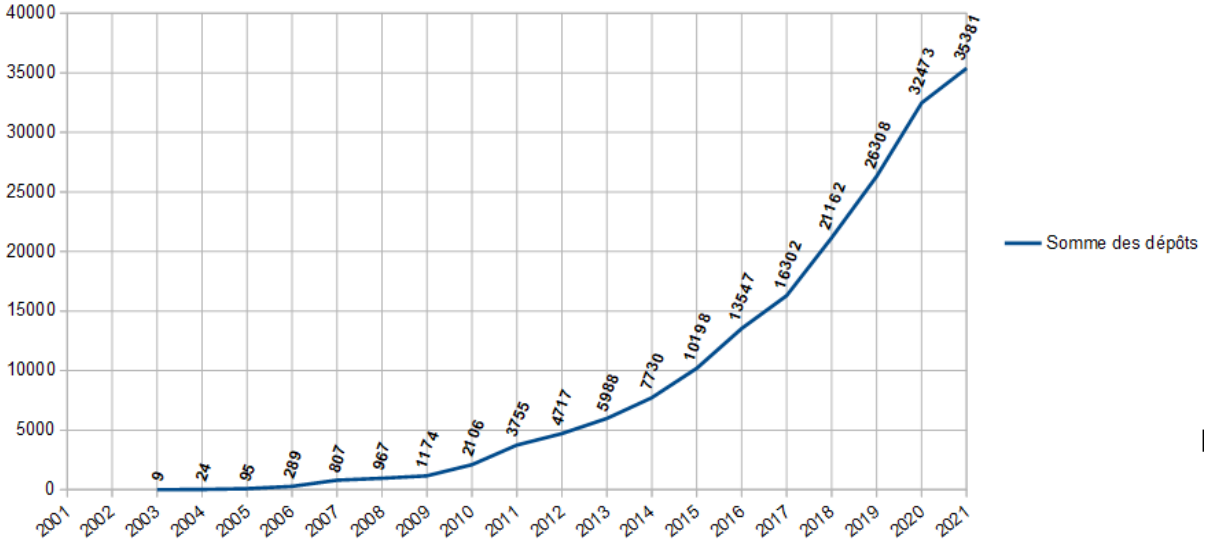

Figure 4: Évolution des dépôts sur HAL des 36 laboratoires (N=35 381)

La « figure 5 » montre le pourcentage de l'évolution des dépôts d'une année à l'autre, par rapport à la somme des dépôts des laboratoires de recherche en SIC sur HAL.

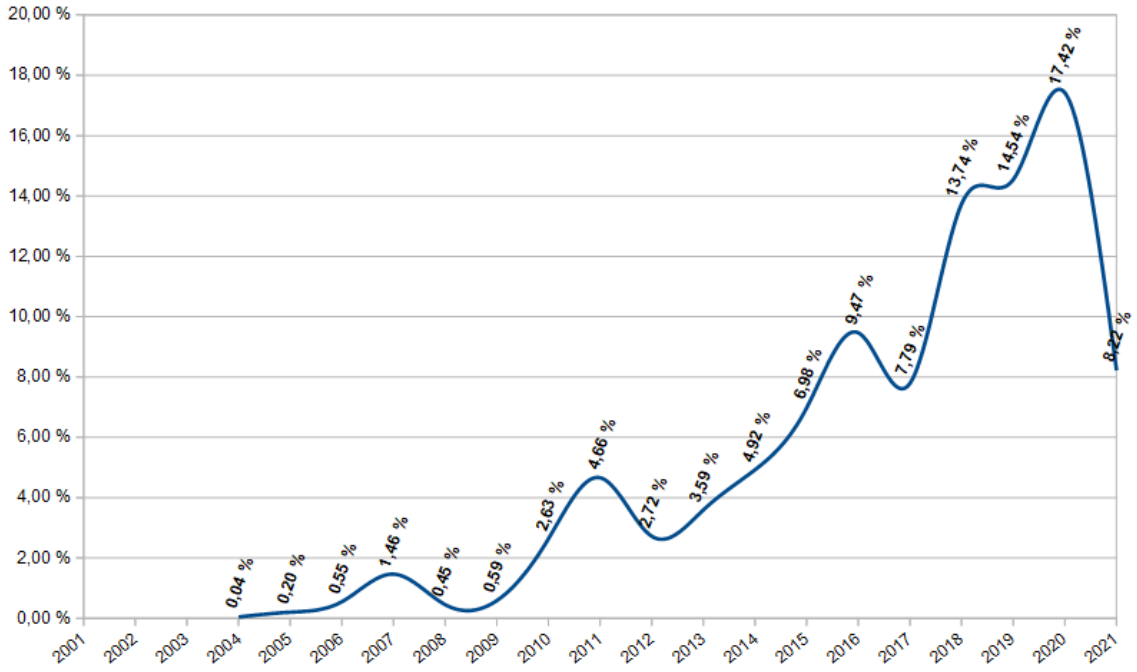

Figure 5: Évolution des dépôts sur HAL des 36 laboratoires en %

2 Typologie des documents :

Selon une requête[52] que nous avons effectué sur l'API de recherche HAL, la plateforme HAL permet le dépôt de 40 types de documents. La somme de notre corpus est de 35 381 dépôts. Nous avons compté 19 types de documents dans les collections des laboratoires de recherche en SIC. Parmi ces 19 types de documents déposés dans les collections des laboratoires de recherche en SIC, nous avons constaté que trois quarts des dépôts (75,83 %) sont constitués de trois types de documents uniquement : des articles, des communications et des chapitres d'ouvrages (Figure 6).

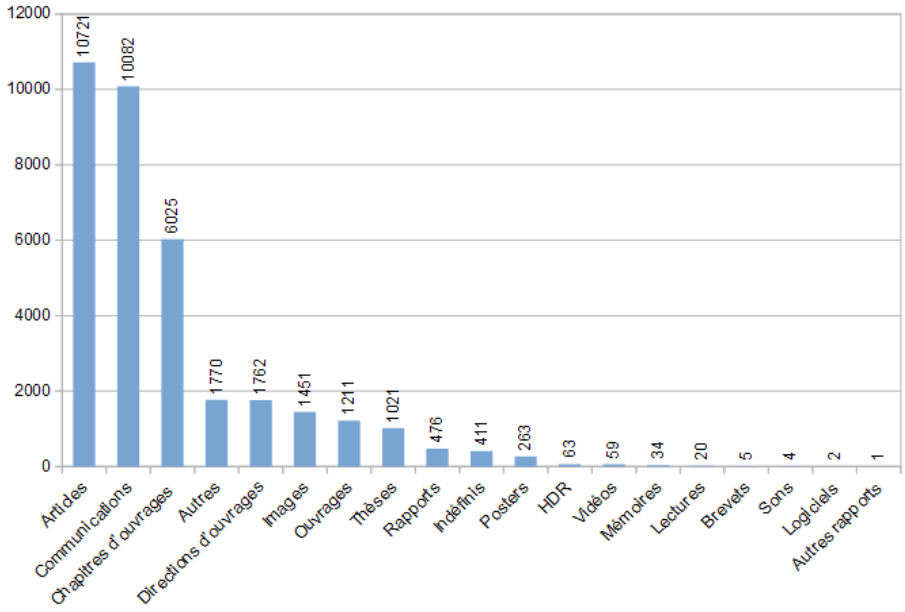

Figure 6: Types de documents déposés dans les collections des laboratoires de recherche en SIC (N=35 381)

Les articles représentent 30 % des dépôts, suivis par des « communications » avec 28 %, puis 17 % des « chapitres d'ouvrages ». Les « autres types » des documents représentent au total 24,17 % des dépôts des laboratoires. Ce taux de 24 % rassemble 16 types de documents dont chacun représente moins de 5 % de l'ensemble des dépôts (Figure 7).

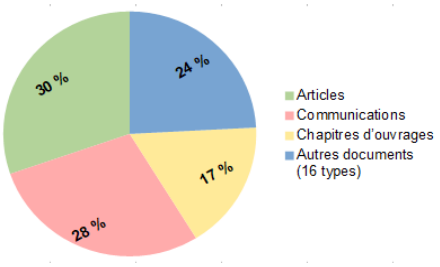

Figure 7: Répartition des types de documents(N=35 381)

52 Requête https://api.archives-ouvertes.fr/search/?q=*:*&facet=true&facet.field=docType_s&rows=0&wt=json

HAL propose un classement des types de documents selon quatre catégories[53] : des publications, documents non encore publiés, travaux universitaires et données de recherche. Parmi les 19 types documents déposés dans les collections HAL des laboratoires de recherche en SIC, nous retrouvons essentiellement des documents de la catégorie « publications » avec 11 types de documents : articles, communications, chapitres d'ouvrages, etc. Ensuite, des documents de la catégorie des « données de recherche » qui comprend 4 types de documents : images, sons, vidéos et logiciels. Puis, 3 types de documents de la catégorie des « travaux universitaires ». Et enfin, un seul document de la catégorie des « documents non encore publiés ». La catégorie « publications » représente une part très importante dans les dépôts des laboratoires de recherche en SIC sur HAL avec un taux de plus de 90 % des dépôts, suivie par les données de recherche qui représente seulement 4,28 % de l'ensemble du corpus (Figure 8).

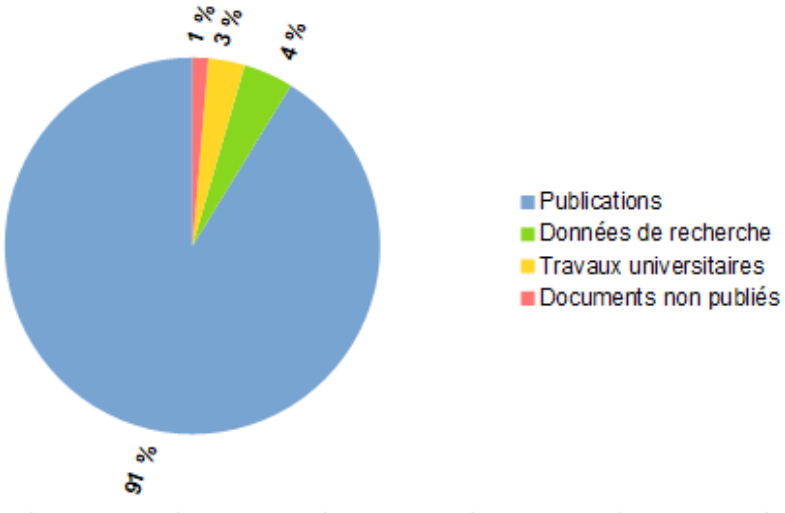

Figure 8: Répartition des dépôts selon la catégorie des documents (N=35 381)

53 HAL. « *Quel type de document choisir – HAL Documentation* ». https://doc.archives-ouvertes.fr/tutoriels/quel-type-de-document-choisir/

3 Part des documents en libre accès :

Les documents « avec un fichier » occupent une place importante dans les collections HAL des laboratoires de recherche en SIC. En effet, le nombre de documents en libre accès s'élève à plus de 10 200 documents par rapport à l'ensemble des dépôts. Ce qui représente donc un tiers des documents de notre corpus (35 381 dépôts). Les notices, quant à elles, représentent près de trois quarts des dépôts des laboratoires de recherche en SIC (Figure 9).

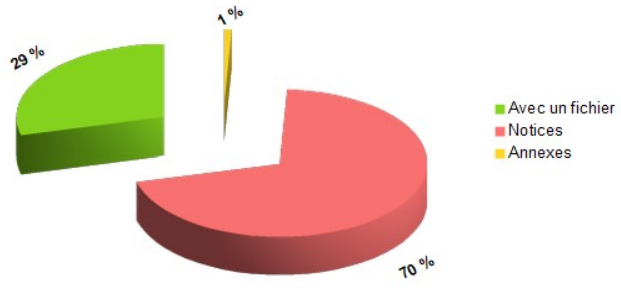

Figure 9: Part des documents en libre accès (N=35 381)

Cette part des documents en libre accès (dépôts avec fichiers) est constituée à la fois des publications (articles, communications, chapitres d'ouvrages,…), des travaux universitaires (thèses et HDR) et des données de recherche (images, sons vidéos et logiciels). Les dépôts relevant des catégories des travaux universitaires et des données de recherche doivent être obligatoirement accompagnés d'un fichier. Ceux là représentent 7 % des dépôts par rapport à l'ensemble des dépôts et 26 % par rapport aux documents avec un fichier. Les dépôts relevant de la catégorie des publications, quant à eux, représentent 22 % de l'ensemble des dépôts et 74 % par rapport aux documents avec un fichier (Figue 10).

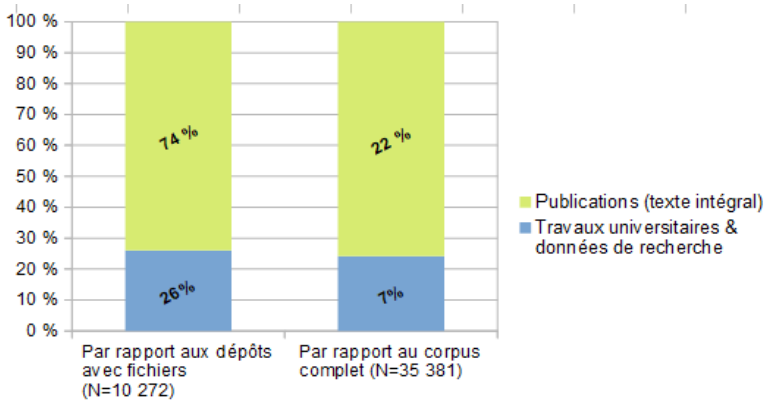

Figure 10: Répartition des documents en libre accès selon leurs catégories

Quand nous observons la distribution de la somme des documents en accès libre par laboratoire, nous nous apercevons que le taux des documents en libre accès diffère d'un laboratoire à l'autre. En effet, nous constatons qu'un tiers des documents en libre accès (34 %) sont déposés par trois laboratoires seulement : CRESAT, CLLE et CREM, qui représentent respectivement 14 %, 11 % et 9 % de la somme des documents en libres accès. Nous retrouvons ensuite le laboratoire GERiiCO, CNE, CERTOP et CEREGE-Lab avec un total de 20 % (respectivement 6 %, 5 %, 5 % et 4 %). Cela signifie que plus de 54 % de la somme des documents en libre accès sont produits par 19 % (7) des laboratoires de recherche sélectionnés (Figue 11).

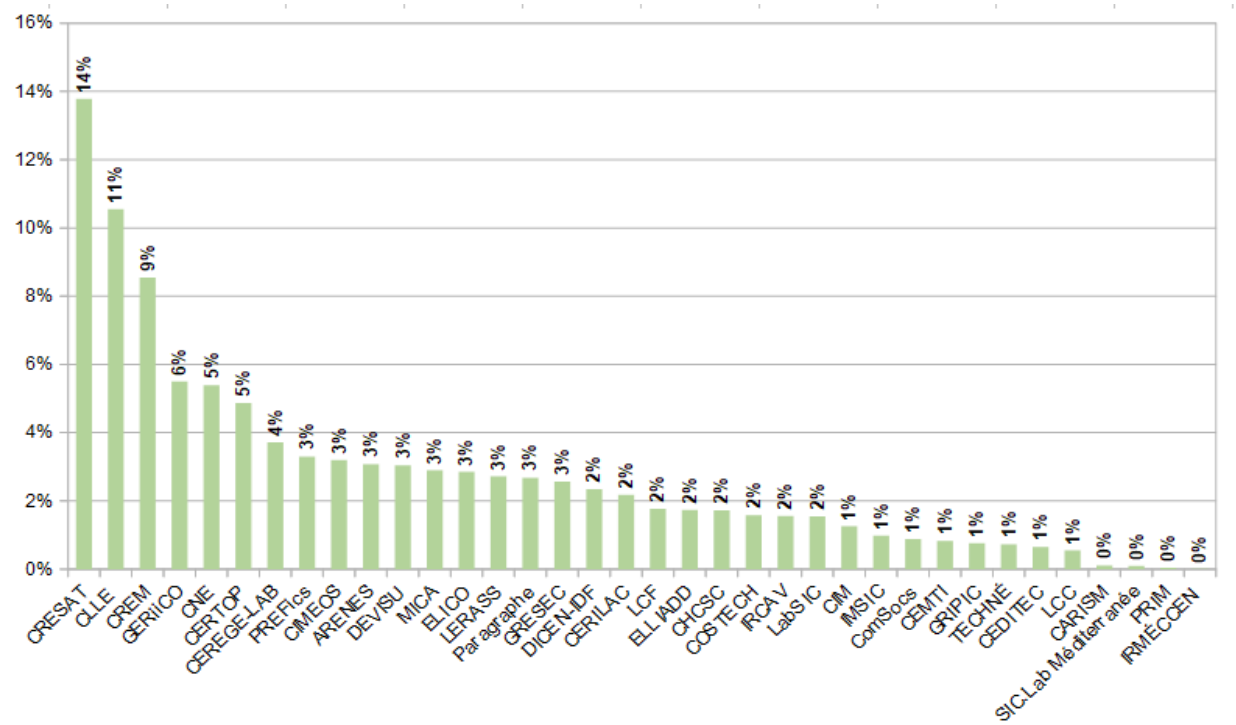

Figure 11: Répartition des documents en libre accès par laboratoire (N=10 272)

Cependant, lorsque nous regardons de plus près les types des dépôts par collection, nous constatons que 44 % des collections (16 collections) comportent au moins 30 % des dépôts avec texte intégral (avec fichier), dont 4 collections sont constituées de 50 à 60 % des dépôts en texte intégral et 2 autres avec 80 % et 82 % des dépôts en texte intégral. Alors que les 20 autres collections (56 %) sont constituées en moyenne de 20 % des dépôts en texte intégral. Ce qui montre que certains laboratoires de recherche en SIC déposent une part importante des documents en libre accès dans leurs collections HAL, tandis que d'autres laboratoires de recherche produisent seulement une partie modeste des documents en libre accès dans HAL (Figure 12).

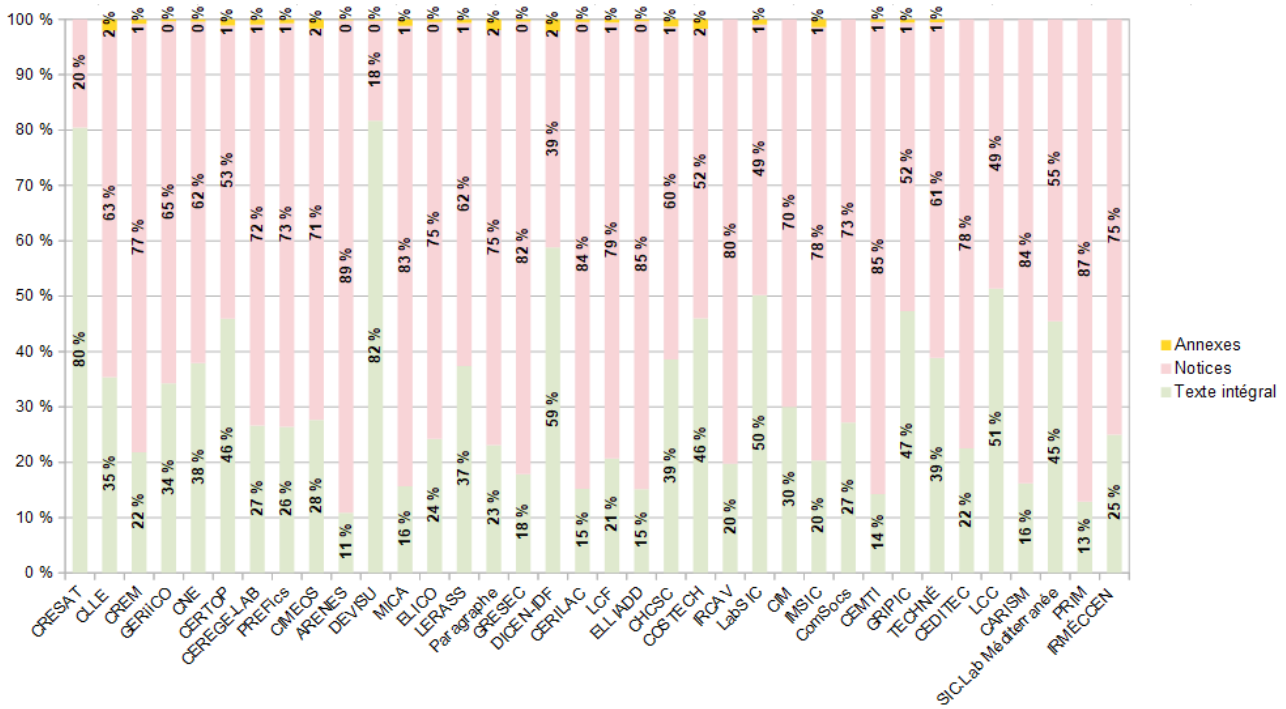

Figure 12: Part des documents en libre accès dans chaque collection (N=35 381)

4 Part des publications en langues étrangères :

Intéressons-nous maintenant à la langue de communication des laboratoires de recherche en SIC sur HAL pour tenter de déterminer le degré d'ouverture internationale de leurs collections. Les dépôts des laboratoires de recherche en SIC sur HAL sont constitués principalement des publications en français, mais aussi des publications en langues étrangères, notamment en anglais. Comme nous pouvons le voir dans la « figure 13 », trois quarts des publications dans les collections HAL des laboratoires de recherche en SIC sont diffusés en français et un quart en langues étrangères dont 19 % des publications sont publiées en anglais et 6 % des publications correspondent à d'autres langues étrangères comme par exemple : allemand, espagnol, italien, etc.

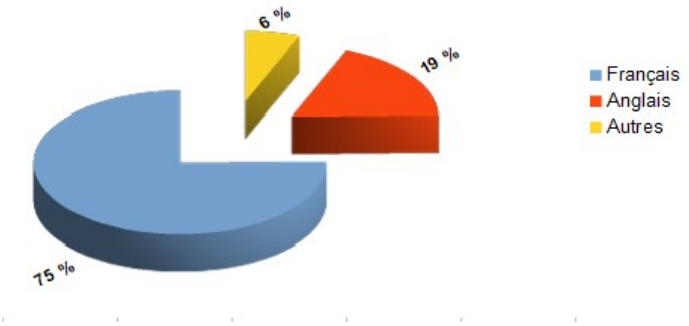

Figure 13: Répartition des dépôts selon la langue de publication (N=35 381)

La répartition des dépôts selon la langue de publication illustrée à la « figure 14 » nous permettrait de nous rendre compte de la part des publications en langues étrangères dans chaque collection. Nous constatons que le taux des dépôts des laboratoires sur HAL varie d'une collection à l'autre. Cependant, nous pouvons voir que la majorité des collections comptent entre plus de 10 % et 40 % de leurs publications en langues étrangères (anglais + autres). Ce qui fait que la moyenne des publications en langues étrangères par collections est de 22 % des dépôts de chaque laboratoire de recherche.

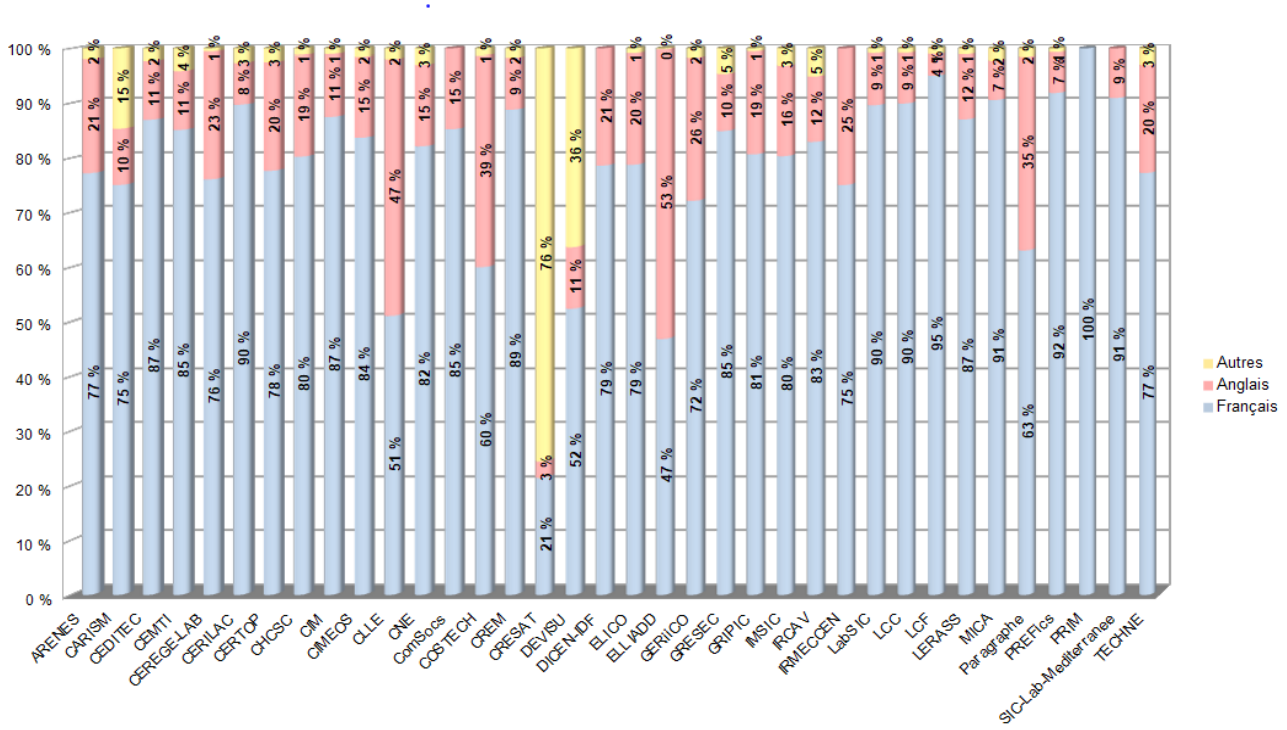

Figure 14: Répartition des dépôts de chaque collection selon la langue de publication (N=35 381)

5 Discussion des résultats :

Notre objectif était d'étudier la présence des laboratoires de recherche en Sciences de l'information et de la communication sur HAL à l'aune de nombre de dépôts, la typologie des documents déposés, la part des documents en libre accès et la langue de publication. Les résultats de cette première approche ont montré une forte présence des laboratoires de recherche en SIC sur HAL. En effet, nos résultats ont apporté un certain nombre d'éclairages sur la production scientifique globale des laboratoires de recherche en SIC dans leurs collections HAL, sur les types des documents qu'ils déposent et sur l'ouverture et l'internationalité de leurs collections, en termes de la part des documents en libre accès et de la langue des publications.

Nous avons constaté que les laboratoires de recherche sont très engagés et contribuent très fortement au dépôt de leurs productions et à l'alimentation des leurs collections dans HAL afin de rendre visible leur production scientifique dans cette plateforme d'archives ouvertes. Nous avons en effet observer une croissance importante de nombre de dépôts, notamment pendant ces cinq dernières années, avec une augmentation moyenne de 12 % des dépôts par année. Cependant, nous avons constaté que certains laboratoires de recherche sont plus présents et plus engagés que d'autres laboratoires. Ce qui signifie que les politiques et les pratiques de dépôts des laboratoires de recherche en SIC sur HAL varient d'un laboratoire à l'autre.

En ce qui concerne la typologie des documents déposés sur HAL par les laboratoires de recherche en SIC, nous avons constaté qu'il y a une large diversité en termes de types des documents déposés dans leurs collections HAL. Cependant, il est important de souligner que les dépôts des articles scientifiques, des communications dans un congrès et des chapitres d'ouvrages occupent une plus grande place dans les collections HAL des laboratoires de recherche en SIC. Cela est valable pour l'ensemble des laboratoires. Annaïg Mahé et Camille Prime-Claverie, dans leur étude de 2013, ont observé quasiment la même pratique chez les sciences de la vie. Elles ont en effet montré que les articles scientifiques constituent une part dominante dans les dépôts sur HAL des sciences de la vie. Cette observation soutient donc nos résultats et montre qu'il y a une certaine similitude entre les différents domaines en matière de types de documents déposés dans HAL.

Ensuite, nous avons constaté que les laboratoires de recherche en SIC ont tendance globalement à déposer des documents avec texte intégral. Nos résultats ont révélé que les documents en libre accès représentent 29 % de la totalité des dépôts des laboratoires de recherche en SIC. La part des documents en libres accès est relativement importante, notamment si l'on pourrait comparer ce taux d'ouverture aux résultats montrés par Joachim Schöpfel dans son étude sur les unités de recherche de Lille. Selon lui, le taux des documents en libre accès sur l'ensemble de HAL est de 25 % et que celui des unités de recherche de l'Université de Lille est de 23 % de leurs dépôts. Ce qui signifie que les laboratoires de recherche en Sciences de l'information et de la communication ont

un taux d'ouverture un peu plus élevé par rapport à l'ensemble des dépôts sur HAL. Cependant, ce taux d'ouverture en matière de libre accès varie également d'un laboratoire à l'autre. Nous avons constaté que tous les laboratoires ne contribuent pas de la même manière au libre accès.

Concernant l'ouverture internationale des collections HAL des laboratoires de recherche en SIC, nous pouvons noter que les laboratoires réservent une place assez importante aux publications en langues étrangères. Un quart de leurs publications dans HAL sont écrites en anglais ou avec une autre langue étrangère. Nous soulignons aussi sur ce point d'ouverture internationale que des différences sont observées entre les laboratoires de recherche. Ils ne réservent pas tous la même place aux publications en langues étrangères, mais les différences restent assez légères sur ce point.

Enfin, nous pouvons relever quelques aspects qui pourraient potentiellement porter des limites à ces résultats. En effet, nous pouvons souligner que notre étude est basée uniquement sur une analyse quantitative des publications des laboratoires de recherche en SIC. Le fait que l'on soit limité par le temps, nous n'avons pas pu intégrer l'approche qualitative. La situation est assez active au sein de certains laboratoires, ce qui veut dire que les données peuvent évoluer à tout moment sur HAL. Un autre point important concernant l'ouverture internationale des collections, nous pouvons souligner que la langue de publication ne suffit pas pour étudier l'internationalisation des collections HAL, car nous avons constaté que la nomenclature HAL ne distingue pas entre les communications dans un congrès national de celles dans un congrès international. Il se pourrait y avoir des publications écrites en français, mais issues d'un évènement ou d'une contribution internationale. Il est donc difficile de déterminer le degré exact d'ouverture internationale des collections HAL uniquement à partir de la langue de publication.

Conclusion

Notre étude sur la présence des laboratoires de recherche en Sciences de l'information et de la communication sur la plateforme HAL n'est qu'une première approche. Cette analyse préliminaire a porté sur 36 laboratoires de recherche en SIC et s'est basée sur des données quantitatives de leurs publications sur HAL. Les résultats ont montré une forte présence des laboratoires de recherche en SIC sur HAL. Cependant, l'étude montre qu'il y a des différences entre les laboratoires de recherche. Tous les laboratoires n'ont pas le même nombre de dépôts dans leurs collections, ne déposent pas tous les mêmes types de documents, n'ont pas la même attitude en ce qui concerne la mise à disposition de leurs productions scientifiques en libre accès ainsi qu'une diversité dans leurs degrés d'ouverture internationale en termes de nombre de publications en langues étrangères. Ce que nous pouvons retenir de ces constats, c'est qu'il n'y a pas « une seule » politique de publication, mais plutôt « des politiques », au pluriel, de publications, même si parfois, nous pouvons constater quelques similitudes et rapprochements entre certains laboratoires de recherche.

La spécificité de cette étude, c'est qu'elle s'est intéressée uniquement aux laboratoires de recherche en Sciences de l'information et de la communication et non pas à des laboratoires de recherche de divers domaines. Aucune étude n'a été menée à ce jour sur la présence des laboratoires en SIC sur HAL. Cette étude n'est qu'un début d'analyse des collections HAL des laboratoires de recherche en SIC et il serait intéressant de poursuivre ce travail pour l'approfondir en analysant séparément les collections des laboratoires pour mieux comprendre la politique de publication sur HAL de chaque laboratoire. Elle serait un point de départ pour les chercheurs souhaitant creuser un peu plus dans ce sujet. Elle permettrait également aux chercheurs et aux laboratoires de recherche en SIC d'avoir une vue globale sur l'ensemble des productions du domaine des SIC dans HAL.

Dans la continuité de cette étude, nous travaillerons sur l'homogénéisation de la nomenclature HAL avec celle du HCERES pour la collection HAL du laboratoire GERiiCO. Il est important que les deux nomenclatures soient homogènes afin d'inciter les chercheurs et les laboratoires à déposer leurs productions dans HAL, car ils sont obligés d'obéir à la nomenclature HCERES dans leurs bilans et dossiers de qualification au CNU. Nous analyserons également les thématiques des publications des laboratoires de recherche en SIC afin d'effectuer un premier pas vers un thésaurus des SIC en français.

Bibliographie

Livres :

- Boure, Robert. « *Les origines des Sciences de l'information et de la communication : Regards croisés* ». Villeneuve d'Ascq : Presses universitaires du Septentrion, 2002.

- Callon, Michel ; Courtial, Jean-Pierre ; Penan, Hervé « *La scientométrie* ». Paris : Presses universitaires de France, 1993.

- Courtial, Jean-Pierre. « *Introduction à la scientométrie : de la bibliométrie à la veille technologique* ». Paris : Anthropos – Economica,1990.

- Gingras, Yves. « *Les dérives de l'évaluation de la recherche : Du bon usage de la bibliométrie* ». Paris : Raisons d'agir éditions, 2014.

- Ibekwe-SanJuan, Fidelia. « *La science de l'information : origines, théories et paradigmes* ». Paris : Hermès science publications-Lavoisier, 2012.

- Lafouge, Thierry ; Le Coadic, Yves-François ; Michel, Christine. « *Éléments de statistique et de mathématique de l'information : infométrie, bibliométrie, médiamétrie, scientométrie, muséométrie, webométrie* ». Villeurbanne : Presses de l'ENSSIB, 2002.

- Le Coadic, Yves-François. « *La science de l'information* ». Paris : Presses Universitaires de France, 1994.

- Noyer, Jean-Max. « *Les sciences de l'information : Bibliométrie, Scientométrie, Infométrie* ». Rennes : Presses universitaires de Rennes,1995

- Olivesi, Stéphane. « *Sciences de l'information et de la communication* ». Presses universitaires de Grenoble, 2006.

- Otlet, Paul « *Traité de documentation : le livre sur le livre : théorie et pratique* ». Préf. de Robert Estivals, av.-pr. de André Canonne. Palais mondial, 1989. https://libstore.ugent.be/fulltxt/handle/1854/5612/Traite_de_documentation_ocr.pdf

- Salaün, Jean-Michel ; Arsenault, Clément. « *Introduction aux sciences de l'information* ». Paris : La Découverte, 2010.

Articles de périodiques :

- Bourdaa, Mélanie, et Lamy, Aurélia. « *Les laboratoires de recherche en Sciences de l'Information et de la Communication* ». 2013, *Revue française des sciences de l'information et de la communication (3)*. https://journals.openedition.org/rfsic/657

- Bourdaa, Mélanie, et Aurélia Lamy. 2014. « *Les enjeux des relations internationales pour les laboratoires de recherche en SIC* ». *Revue française des sciences de l'information et de la communication (4)*. https://journals.openedition.org/rfsic/805

- Fondin Hubert, « *La science de l'information : posture épistémologique et spécificité disciplinaire* », *Documentaliste-Sciences de l'Information*, 2001/2 (Vol. 38), p. 112-122. https://www.cairn.info/revue-documentaliste-sciences-de-l-information-2001-2-page-112.htm

- Fondin Hubert, « *La Science de l'information ou le poids de l'histoire [1]* », *Les Enjeux de l'information et de la communication*, 2005/1 (Volume 2005), p. 35-54. https://www.cairn.info/revue-les-enjeux-de-l-information-et-de-la-communication-2005-1-page-35.htm

- Mahé, Annaïg, et Camille Prime-Claverie. « *Qui dépose quoi sur Hal-SHS ? Pratiques de dépôts en libre accès en sciences humaines et sociales* ». 2017. *Revue française des sciences de l'information et de la communication (11)*. https://journals.openedition.org/rfsic/3315

- Prime-Claverie, et Camille Mahé Annaïg, « *Sites de dépôt en libre accès et formes de médiations : quelles évolutions?* ». *Joumana Boustany éd., La médiation numérique : renouvellement et diversification des pratiques*. 2013, p. 125-139. https://www.cairn.info/la-mediation-numerique-renouvellement--9782804182274-page-125.htm

- Schöpfel Joachim, « *L'usage de la plateforme HAL par des unités de recherche. Le cas de l'Université de Lille* », I2D - Information, données & documents, 2020/3 (n° 3), p. 167-198. https://www.cairn.info/revue-i2d-information-donnees-et-documents-2020-3-page-167.htm

- Schöpfel, Joachim, Hélène Prost, Amel Fraisse, et Stéphane Chaudiron. « *Valoriser les publications d'un laboratoire universitaire dans l'environnement de la science ouverte* ». *HAL archive ouverte* 2018. https://hal.archives-ouvertes.fr/hal-01940352

- Tomic Yves, « *De l'usage des API. Les API de l'Abes* », *Documentaliste-Sciences de l'Information*, 2014/3 (Vol. 51), p. 17-18. https://www.cairn.info/revue-documentaliste-sciences-de-l-information-2014-3-page-17.htm

Articles sur Internet :

- Emile Gayoso. « *La diffusion sur Hal, Academia et ResearchGate des articles de recherche des revues françaises de Sciences Humaines et Sociales* », 2020. Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation. Consulté 12 mai 2021. https://www.enseignementsup-recherche.gouv.fr/cid149392/www.enseignementsup-recherche.gouv.fr/cid149392/la-diffusion-sur-hal-academia-et-researchgate-des-articles-de-recherche-des-revues-francaises-de-s.h.s.html

Autres :

- GIS Réseau Urfist « *HAL/LO – Valorisation Sur HAL de La Production Des Laboratoires Dans l'environnement de La Science Ouverte* ». Consulté 1 juin 2021. https://gis-reseau-urfist.fr/hal-lo-valorisation-sur-hal-de-la-production-des-laboratoires-dans-lenvironnement-de-la-science-ouverte/

- HAL. « *Quel type de document choisir – HAL Documentation* ». Consulté 2 juin 2021. https://doc.archives-ouvertes.fr/tutoriels/quel-type-de-document-choisir/

Annexes

Constitution de la liste des laboratoires de recherche en SIC :

- CPDirSIC. « *Membres de la CPDirSIC* ». Consulté 5 décembre 2020. http://cpdirsic.fr/membres-de-la-cpdirsic/

- Ministère français de l'Enseignement supérieur, de la Recherche et de l'Innovation « *Liste des Unités de Recherche, de la discipline : Sciences de l'information et de la communication* ». Consulté 5 décembre 2020. https://appliweb.dgri.education.fr/annuaire/ListeEntite.jsp?entite=ur&sd=22&prov=MotCle

- RNSR : https://appliweb.dgri.education.fr/rnsr/ChoixCriteres.jsp?PUBLIC=OK

- ScanR : https://scanr.enseignementsup-recherche.gouv.fr/

- Hcéres : https://www.hceres.fr/fr/rechercher-une-publication?key=&f%5B0%5D=themes_publications%3A43

- AURéHAL : https://aurehal.archives-ouvertes.fr/structure

Documentation API de recherche HAL :

- CCSD « *API HAL | API de recherche HAL* ». Consulté 2 juin 2021. https://api.archives-ouvertes.fr/docs/search

Requêtes utilisées :

- Extraction de nombre de dépôts sur un intervalle d'années : https://api.archives-ouvertes.fr/search/GERIICO/?q=*:*&wt=csv&rows=5000&fq=submittedDateY_i:[2000%20TO%202021]

- Des dépôts avec fichiers : https://api.archives-ouvertes.fr/search/GERIICO/?q=*:*&wt=csv&rows=5000&fq=submitType_s:file

- Extraction de type de documents, type de dépôts, date de dépôt et la langue de publication : https://api.archives-ouvertes.fr/search/GERIICO/?q=*:*&wt=csv&rows=5000&indent=true&facet=true&fl=docid,halId_s,version_i,docType_s,submitType_s,submittedDateY_i,language_s

- Extraction de la version du document, type de documents, titre, nom(s) de d'auteur(s), acronyme laboratoire(s), acronyme institution(s), citation abrégé, titre de revue, URI, type de dépôts, année de dépôt, mots-clés, langue de document, nom du contributeur (déposant), regroupement de laboratoires et regroupement d'institutions (acronyme) : https://api.archives-ouvertes.fr/search/GERIICO/?q=*:*&wt=csv&rows=5000&indent=true&facet=true&fl=docid,halId_s,version_i,docType_s,title_s,authFullName_s,labStructAcronym_s,instStructAcronym_s,citationRef_s,journalTitle_s,uri_s,submitType_s,submittedDateY_i,keyword_s,language_s,contributorFullName_s,rgrpLabStructAcronym_s,rgrpInstStructName_s

- Idem, mais en utilisant le code de structure : https://api.archives-ouvertes.fr/search/?q=structId_i:39707&wt=csv&rows=5000&indent=true&facet=true&fl=docid,halId_s,version_i,docType_s,title_s,authFullName_s,labStructAcronym_s,instStructAcronym_s,citationRef_s,journalTitle_s,uri_s,submitType_s,submittedDateY_i,keyword_s,language_s,contributorFullName_s,rgrpLabStructAcronym_s,rgrpInstStructName_s

Résumé :

Cette étude fournit une première approche d'évaluation de la production scientifique des laboratoires de recherche en Sciences de l'information et de la communication (SIC) sur la plateforme d'archives ouvertes HAL. Elle porte sur les publications de 36 laboratoires de recherche en SIC dans HAL et présente les résultats d'une analyse bibliométrique réalisée à partir des données extraites via HAL. L'objectif de cette recherche est d'analyser la présence des laboratoires de recherche en SIC sur HAL en fonction d'un certain nombre de variables : le nombre de dépôts, la langue de publication, la typologie des documents et la part des documents en libre accès. Les résultats de l'étude ont montré une présence importante des laboratoires en SIC sur HAL, mais nous avons également constaté des différences entre les laboratoires de recherche concernant leur nombre dépôts, leur ouverture en termes de libre accès ainsi que l'internationalisation de leurs collections. Chaque laboratoire a sa propre politique de publication sur HAL.

Mots-clés :

Science ouverte, libre accès, archive ouverte, HAL, API, analyse bibliométrique, production scientifique, laboratoire de recherche en Sciences de l'information et de la communication.