# Valoriser les publications d'un laboratoire universitaire dans l'environnement de la science ouverte

Joachim Schöpfel, Hélène Prost, Amel Fraisse, Stéphane Chaudiron

# Valoriser les publications d'un laboratoire universitaire dans l'environnement de la science ouverte

## Retour d'expérience de la collection GERiiCO sur HAL

Joachim Schöpfel est enseignant-chercheur (MCF-HC) en Sciences de l'information et de la communication à l'Université de Lille et membre du laboratoire GERiiCO. Après avoir dirigé l'UFR IDIST de 2009 à 2012 et l'Atelier National de Reproduction des Thèses de 2012 à 2017, il travaille aujourd'hui comme consultant indépendant dans le domaine de l'information scientifique. Ses intérêts scientifiques : le libre accès à l'information, les données de la recherche, les thèses et mémoires, la science ouverte.
ORCID 0000-0002-4000-807X joachim.schopfel@univ-lille.fr

Hélène Prost est ingénieur d'études à l'Institut de l'Information Scientifique et Technique (INIST-CNRS) et membre associé au laboratoire GERiiCO de l'Université de Lille. Elle participe à différents projets de recherche relatifs à l'évaluation des collections, la fourniture de documents et l'analyse des usages, ainsi qu'à la littérature grise, les données de la recherche et le libre accès; elle est référente publications de la collection GERiiCO sur HAL. Elle est l'auteur de plusieurs publications.
ORCID 0000-0002-7982-2765 helene.prost@inist.fr

Amel Fraisse est Maître de conférences en Sciences de l'information et de la communication à l'Université de Lille et membre du laboratoire GERiiCO. Ses travaux de recherche s'inscrivent dans le domaine du Traitement Automatique de la Langue et plus particulièrement la collecte, le traitement et la diffusion d'information multilingues.
ORCID 0000-0002-8693-8862 amel.fraisse@univ-lille.fr

Stéphane Chaudiron est professeur en Sciences de l'information et de la communication à l'Université de Lille et directeur du laboratoire GERiiCO. Il est co-directeur de la revue *Etudes de Communication,* président du comité scientifique de la revue *I2D - Information, Données, Document,* responsable pédagogique du parcours de Master VeCIS (Veille et Communication de l'Information Stratégique), membre du conseil d'UFR, membre de l'Association for information science and technology (ASIST), membre du chapitre français de l'International society for knowledge organization (ISKO) dont il été président de 2005 à 2009, co-animateur du groupe de travail « Veille et analytique » du Groupement français de l'industrie de l'information » (GFII) et membre de l'Association des professionnels de l'information et de la documentation (ADBS). Il a été chargé de mission "politique éditoriale" auprès de la Présidence de l'Université de Lille 3 et membre du conseil d'administration de cette université.
stephane.chaudiron@univ-lille.fr

## Résumé

La question de la diffusion des résultats de la recherche et, en particulier, le libre accès aux publications des chercheurs est au cœur de la politique pour la science ouverte. Comment

peut se positionner un laboratoire de recherche universitaire ? Comment peut se traduire la politique pour la science ouverte sur le terrain d'un campus universitaire ? Sous forme d'un retour d'expérience, notre étude analyse la mise en place de la collection du laboratoire GERiiCO de l'Université de Lille sur l'archive ouverte nationale HAL. L'objectif de l'initiative est double : d'une part, assurer une visibilité maximale et un impact au-delà de la communauté disciplinaire, à travers des médias sociaux et le référencement des moteurs de recherche ; d'autre part, contribuer à l'évaluation de la production scientifique du laboratoire. Nous présentons les ressources mobilisées et les actions mises en oeuvre, analysons les résultats en termes de dépôts, d'usage et de services, et évoquons les facteurs de succès, les problèmes rencontrés et quelques perspectives pour le futur développement. En particulier, nous comparons le contenu de la collection HAL avec les résultats de la base de données scientométrique d'Elsevier (Scopus) et du moteur de recherche Google Scholar, et nous montrons le potentiel de la collection pour visualiser les relations au sein du laboratoire (analyse de réseaux) et son rayonnement international.

### Mots clés

Laboratoire universitaire, production scientifique, publications, archive ouverte, évaluation, valorisation, HAL, science ouverte

## Introduction

Le 4 juillet 2018, lors de la conférence annuelle de la Ligue des Bibliothèques Européennes de Rechercher à Lille, la Ministre de l'Enseignement Supérieur, de la Recherche et de l'Innovation, Frédérique Vidal, a annoncé un Plan national pour la science ouverte. L'ambition est de rendre "les résultats de la recherche scientifique ouverts à tous, sans entrave, sans délai, sans paiement." Le premier axe du Plan est de généraliser l'accès ouvert aux publications, de rendre obligatoire la publication en accès ouvert des articles et livres issus de recherches financées par appel d'offres sur fonds publics et de soutenir l'archive ouverte nationale HAL.

Lancée en 2002 par le Centre pour la Communication Scientifique Directe (CCSD) du CNRS, l'archive pluridisciplinaire HAL (= Hyper articles en ligne) est devenue au fil des ans l'une des plus importantes plateformes de la "voie verte" du libre accès à l'information scientifique. "Voie verte" veut dire: l'auto-archivage des publications scientifiques par les auteurs eux-mêmes, sur une plateforme dédiée (Harnad et al., 2004). C'est, avec les revues en libre accès ("voie dorée"), l'une des deux principales stratégies pour atteindre, selon les mots de la Ministre, "à terme 100% de publications scientifiques françaises en accès ouvert".

Cette stratégie de la "voie verte" s'appuie avant tout sur deux acteurs : l'auteur, dans la mesure où il détient les droits intellectuels pour effectuer le dépôt de ses propres publications (article, chapitre, communication, mémoire etc.), et son institution (organisme de recherche, université, école etc.), dans la mesure où elle a la possibilité d'inciter voire d'imposer l'auto-archivage (Thirion & Rentier, 2014) et où elle dispose également des ressources et de la légitimité pour une archive institutionnelle (Lynch, 2003). En revanche, un troisième acteur sur l'échiquier de l'Enseignement Supérieur et de la Recherche (ESR) est plus ou moins absent et peu visible dans cette stratégie : le laboratoire de recherche.

Or, les chercheurs de l'ESR français sont tous rattachés à des laboratoires de recherche, des unités constituées affiliées à une université et/ou à un ou plusieurs organisme(s) de recherche. Le laboratoire structure le cadre de travail des chercheurs ; les projets de recherche sont organisés autour des laboratoires, et c'est dans le cadre de leur laboratoire que les chercheurs

sont évalués par le Haut Conseil de l'évaluation de la recherche et de l'enseignement supérieur (HCERES).
A partir de la collection du laboratoire GERiiCO de l'Université de Lille, nous interrogeons le rôle potentiel d'un laboratoire universitaire pour une stratégie de la "voie verte", dans l'environnement d'une politique de science ouverte. Quel est l'intérêt pour un laboratoire, en termes d'impact et de visibilité ? Quels sont les facteurs-clés pour réussir une telle initiative, et quels sont les verrous ? Nous présentons un premier bilan, suivi de quelques perspectives pour un futur développement d'une telle collection.

## Un nouvel écosystème

Comme un rapport récent de l'OCDE le souligne, ce terme de "science ouverte" a beaucoup de significations différentes ; dans un sens très large, il désigne tous les efforts pour rendre le processus scientifique plus ouvert et inclusif pour l'ensemble des acteurs concernés, tant au sein des communautés scientifiques, qu'à l'extérieur du milieu de la recherche (Dai et al. 2018). Malgré les différences d'approches et d'interprétations disciplinaires, il y a consensus qu'une telle politique de science ouverte impacte l'ensemble des acteurs et des dimensions de la recherche, tels que les mécanismes de financement, les modes d'évaluation (*peer review*, indicateurs etc.), les infrastructures scientifiques, le transfert des connaissances, la propriété intellectuelle et la gestion des données de la recherche, mais aussi les nouveaux modèles de "crowdsourcing" et de la science citoyenne.
La question de la diffusion des résultats de la recherche et, en particulier, le libre accès aux publications des chercheurs est au coeur de la Science ouverte. Les enjeux sont connus : rendre les résultats scientifiques accessibles sans délais et sans restrictions, aux autres chercheurs, aux acteurs économiques et aux citoyens, augmenter l'impact et la visibilité des structures scientifiques, maîtriser les coûts de l'information scientifique. Les différentes options du libre accès sont également connues, dont notamment la publication dans des revues en libre accès ("voie dorée") et la diffusion via des archives ouvertes ("voie verte"). Cette dernière option est facilitée par la Loi pour une République numérique de 2016 et la création d'un droit d'exploitation secondaire pour les chercheurs français (article 30, cf. CNRS-DIST 2016).
Comme son modèle américain, le service d'e-prints arXiv[1], l'archive ouverte HAL[2] a été conçue sur le principe de la communication directe entre chercheurs, pour faciliter et accélérer l'échange d'articles, avant même leur publication dans une revue. Sa mission est le dépôt et la diffusion non seulement d'articles scientifiques de niveau recherche, publiés ou non, mais aussi de thèses, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Avec l'évolution du mouvement du libre accès (cf. Suber, 2012) et des archives institutionnelles, puis, notamment après la signature à l'Académie des Sciences le 2 avril 2013, de la "Convention de partenariat en faveur des archives ouvertes et de la plateforme mutualisée HAL" entre les établissements et organismes de l'ESR français, HAL est devenue une sorte d'archive institutionnelle nationale, une "infrastructure nationale mutualisée hébergeant des archives institutionnelles ou vers laquelle les autres archives institutionnelles sont fermement invitées à déverser leur contenu" (Bauin, 2014, p.3).
Cette position singulière au cœur de l'écosystème des archives ouvertes et de la stratégie "voie verte" en France a été confirmée récemment par le Plan national pour la science ouverte; ce Plan rend obligatoire la publication en accès ouvert des articles et livres issus de

[1] https://arxiv.org/
[2] https://hal.archives-ouvertes.fr/

recherches financées par appel d'offres sur fonds publics, il annonce la création d'un fond pour la science ouverte et s'engage à soutenir l'archive ouverte nationale HAL; il déclare également vouloir simplifier le dépôt par les chercheurs qui publient en accès ouvert sur d'autres plateformes dans le monde (MESRI, 2018).
Autour de l'archive ouverte HAL s'est construite une offre de services importante et variée, faite de portails et de collections, de référentiels, d'applications, d'un serveur de thèses de doctorat et d'habilitations (TEL), d'une archive ouverte de photographies et d'images scientifiques (MédiHAL), d'une plateforme pour organiser des événements scientifiques (sciencesconf.org) et d'une plateforme pour réaliser des "épirevues" à moindre coût afin de mettre en œuvre le libre accès aux versions électroniques des articles (Episciences). Une partie des universités et organismes scientifiques ont fait le choix de créer leur archive institutionnelle sur HAL, en agrégeant les dépôts de leurs communautés scientifiques à partir des affiliations. Il s'agit par exemple de Sorbonne Université, de l'Ecole Normale Supérieure, de l'Ecole Centrale de Paris, d'Aix-Marseille Université, de l'Université de Toulouse Paul Sabatier, de l'Université de Lorraine, de l'Ecole Polytechnique et de l'Université de Nantes. Parmi les organismes, on trouve l'INRIA, l'INSERM, l'IRD et le CIRAD, mais aussi le CEA, l'Institut Pasteur ou l'Observatoire de Paris.
GERiiCO fait partie des laboratoires, instituts et autres structures de recherche qui ont décidé la création d'une collection sur HAL, proposant un accès unique aux publications de leurs chercheurs et quelques services annexes. A titre d'exemple, des 66 laboratoires de l'Université de Lille, pour 59 d'entre eux, on trouve des dépôts sur HAL (en tout, 7 582 documents, plus 29 363 notices) mais seulement 20 laboratoires ont créé leur propre collection, pour rendre visible leur production; leurs dépôts de documents représentent un peu plus de 60% des documents de tous les laboratoires.

## Méthodologie

Sous forme d'un retour d'expérience, nous présentons la mise en place de la collection du laboratoire GERiiCO de l'Université de Lille sur l'archive ouverte nationale HAL. Les ressources mobilisées et les actions mises en œuvre font l'objet d'une description succincte, suivie d'une analyse des résultats en termes de dépôts, d'usage et de services et d'une synthèse des facteurs de succès et des problèmes rencontrés. L'étude terminera par quelques perspectives pour le futur développement d'une telle collection de laboratoire.
Les analyses statistiques du contenu et de l'usage de la collection ont été menées en avril et mai 2018, grâce aux outils de la plateforme HAL et en exportant les résultats sous forme de tableaux Excel pour d'autres analyses plus poussées. L'analyse scientométrique des auteurs et co-auteurs a été menée en juillet 2018 avec Gephi, à partir des données extraites en avril et mai. La comparaison scientométrique des publications de GERiiCO dans Scopus, le Web of Science et Google Scholar a été faite en juillet 2018.

## La collection GERiiCO sur HAL

Le laboratoire GERiiCO[3] est un pôle de recherche à vocation internationale en sciences de l'information et de la communication de la région Hauts-de-France. Créé en 2006 comme une équipe d'accueil de l'Université de Lille 3, GERiiCO compte au 1er janvier 2018 35 enseignants-chercheurs titulaires, trois professeurs émérites et 38 doctorants. L'orientation majeure de GERiiCO est l'analyse des pratiques, des processus et des dispositifs info-

[3] Groupe d'Études et de Recherche Interdisciplinaire en Information et COmmunication cf. https://geriico-recherche.univ-lille3.fr/

communicationnels saisis dans leurs dimensions langagières et discursives, technologiques et symboliques.
Les premiers dépôts de la part d'auteurs de GERiiCO remontent à 2003. Pendant plusieurs années, il n'y a pas eu de position officielle ou d'incitation de GERiiCO concernant le libre accès ou HAL ; tous les ans ont été déposés entre 20 et 40 documents ou références sur HAL, par les auteurs eux-mêmes. La situation évolue après la dernière campagne d'évaluation (AERES) du laboratoire en 2014 lorsqu'il est décidé de créer des notices sur HAL pour la production de GERiiCO, à partir des listes bibliographiques 2008-2012 des membres titulaires. En parallèle, une collection GERiiCO est officiellement crée sur HAL, sous la responsabilité d'un administrateur, avec l'aide du SCD. L'objectif de la mise en place de cette collection est double : d'une part, assurer une visibilité maximale et un impact au-delà de la communauté disciplinaire, à travers des médias sociaux et le référencement des moteurs de recherche; d'autre part, pallier l'absence d'un système d'information recherche pour la gestion des publications au sein du laboratoire (cf. figure 1).


*Figure 1. Page d'accueil de la collection GERiiCO sur HAL (capture d'écran, 26 juillet 2018)*

A l'occasion du bilan de l'activité de GERiiCO pour la nouvelle campagne d'évaluation 2019 (HCERES), les références des listes bibliographiques 2013-2017 des chercheurs de GERiiCO ont servi pour la création des notices sur HAL. En même temps, un compte HAL a été créé pour chaque chercheur n'ayant pas un tel compte au préalable. La saisie sur HAL a été réalisée par deux membres de GERiiCO et une vacataire du SCD, après dédoublonnage des listes, et uniquement pour les publications pas encore signalées sur HAL. Une partie des références ont été importées en format BibTex, d'autres ont été générées à partir de leur DOI, le reste a été saisi manuellement.
La saisie de la production 2013-2017 est terminée ; elle a été accompagnée d'un débat au sein du laboratoire, sur l'intérêt d'une telle action et sur la qualité du résultat.

## Documents et notices

A l'issue de la saisie des références du bilan 2013-2017, la collection GERiiCO sur HAL comptait 1 253 notices, dont 413 avec le texte intégral (33%). Elle est composée de treize catégories de documents, avant tout de communications dans un congrès (36%), d'articles de revues (32%) et de chapitres d'ouvrages (15%). Mais on trouve également des rapports, thèses, HDR et posters (tableau 1).

| **Type document** | **Nombre** |
|---|---|
| Communications | 450 |
| Articles | 403 |
| Chapitres d'ouvrages | 184 |
| Directions d'ouvrages | 60 |
| Autres | 41 |
| Rapports | 35 |
| Indéfinis | 28 |
| Thèses de doctorat | 19 |
| Posters | 14 |
| Ouvrages | 12 |
| Mémoires | 3 |
| HDR | 2 |
| Brevets | 2 |
| **TOTAL** | 1 253 |

*Tableau 1: Typologie des documents de la collection GERiiCO sur HAL (N=1 253)*

Les publications les plus anciennes de la collection datent de 1993 à 1999. En fait, 4% correspondent à des publications avant la création de GERiiCO en 2006. Mais il n'y a pas eu de projet pour signaler d'une manière rétrospective la production issue des deux unités de recherche dont la fusion est à l'origine de GERiiCO.
Trois quarts des publications de GERiiCO sont écrites en français, le reste est en anglais, plus quelques textes en italien, polonais, espagnol et allemand (figure 2).

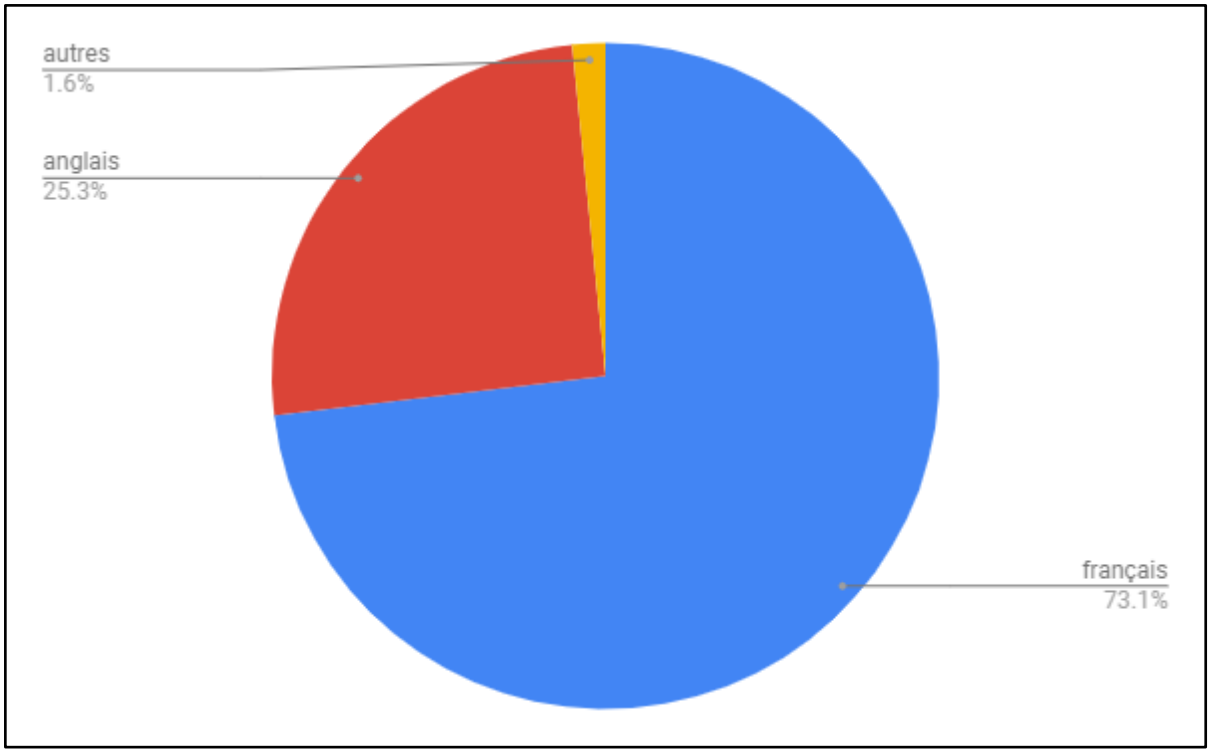


*Figure 2. La langue des publications de la collection GERiiCO sur HAL (n=1 253)*

Nous reviendrons plus loin sur l'aspect international de la collection.

## Exhaustivité et visibilité

L'objectif de la saisie des références à partir des listes bibliographiques du bilan pour le HCERES, en complémentarité de l'auto-archivage par les auteurs est d'obtenir une vitrine exhaustive et fiable de la production des chercheurs de GERiiCO. Pour deux raisons : les grandes bases de données scientométriques (Web of Sciences, Scopus) ne contiennent qu'une petite partie des publications françaises, notamment en SHS (Schöpfel & Prost 2009) ; et le référencement par les outils de recherche comme Google Scholar, Microsoft Academic, Bielefeld Academic Search Engine ou Dimensions reste opaque et manque de fiabilité.
La comparaison directe des publications de GERiiCO sur HAL, dans Google Scholar et dans Scopus confirme ce constat. En effet, l'interrogation de Scopus sur le champ affiliation fournit 74 notices liées à GERiiCO, soit 5,9% de la collection. Les années de publication de ces notices s'étendent de 2006 à 2018, mais seule l'année 2016 atteint 10 notices. Quant à l'interrogation de Google Scholar, on obtient 1 020 réponses dont une partie est du bruit - soit des pages web sans caractère de publications (appel à communication, annonce de conférence…), soit des publications qui ne font pas partie de la production de GERiiCO (communications d'une conférence organisée par GERiiCO…).
Figure 3 illustre les effets des deux montées en charge de saisie dans HAL en 2016 et 2018 sur l'exhaustivité de la collection. Le nombre de notices présentes dans Google Scholar est plus important que celui détenu dans HAL seulement pour l'année de publication 2012, année où le laboratoire GERiiCO a accueilli le colloque *Communiquer dans un monde de normes*, mais aussi pour 2017 et 2018, car la visibilité dans Google Scholar est effective sitôt la saisie sur Internet. Mais encore une fois, une partie des références de Google Scholar sont des erreurs, dans la mesure où elles ne correspondent pas aux publications des chercheurs de GERiiCO.

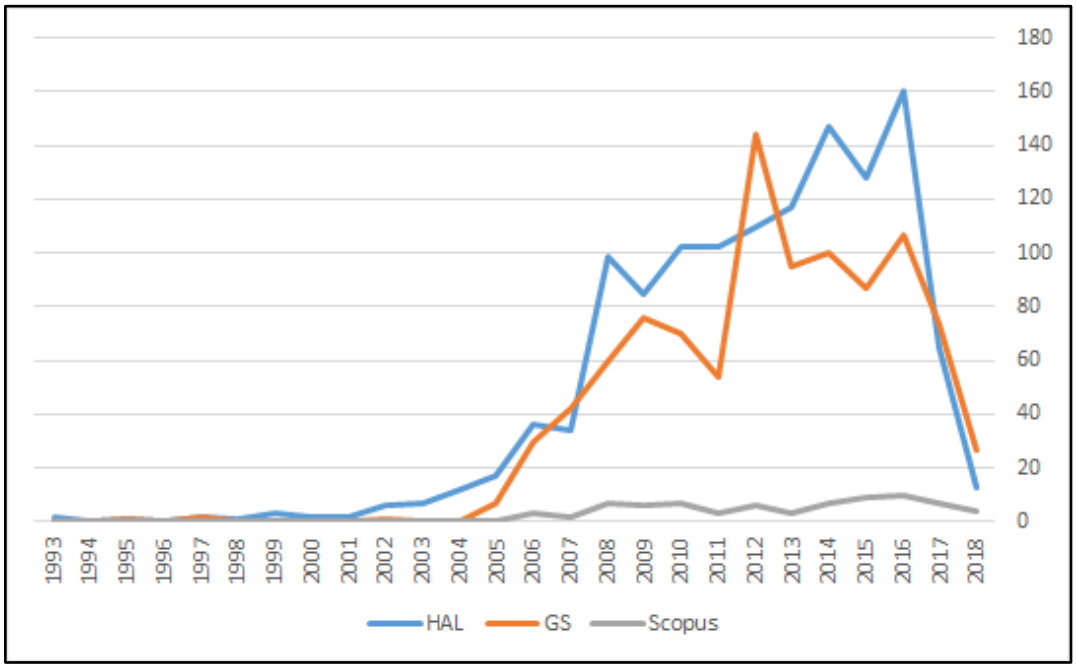


*Figure 3. Comparaison de la collection HAL avec Google Scholar et Scopus (29 juillet 2018)*

Quant à l'exhaustivité de la collection sur HAL, le seul bémol est que par rapport aux listes exhaustives du bilan pour le HCERES, une petite partie des publications ne correspond pas aux critères initiaux de HAL. En effet, au départ, il n'était pas possible d'intégrer certains types de documents dans HAL, tels que les articles de blog, les notes de lecture ou les traductions; c'est désormais possible, mais la saisie rétrospective de ces notices reste à faire Autrement dit, certaines de ces références ne figurent pas dans HAL mais sont potentiellement visibles sur Internet, à condition d'être signées avec l'affiliation GERiiCO.

## Consultations et téléchargements

Dans le cadre de ce travail de recherche, nous nous sommes basés uniquement sur le nombre de consultations des notices et de téléchargements des fichiers comme indicateur d'influence scientifique.

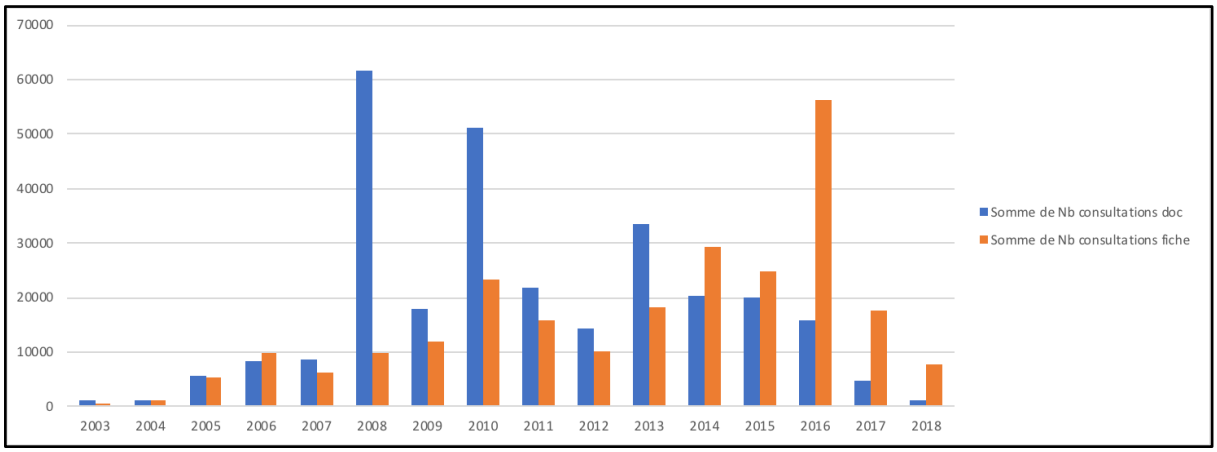


*Figure 4. Consultations et téléchargements de la collection GERiiCO par année de dépôt (2003-2018)*

La figure 4 présente le total des consultations et de téléchargements par document et par référence déposés entre 2003 et 2018. Nous pouvons constater que des documents déposés en 2008 et 2010 ont été marqués par un nombre de consultations très importants. On remarque également un nombre de consultation de références important pour 2016, qui coïncide avec la date de dépôt sur HAL d'une grande partie de la collection GERiiCO.
La figure 5 montre l'évolution du nombre total de documents déposés dans HAL. Ce nombre a évolué de façon constante entre les années 2003 et 2015. Le nombre de documents déposés sur cette période est passé de 0 à 50 documents avec une moyenne de 4-5 dépôts par année. Un nombre de dépôt important (450 documents) a été réalisé en 2016. Ceci s'explique par les actions internes menées au sein du laboratoire pour inciter les chercheurs à déposer leurs documents sur la plateforme HAL.

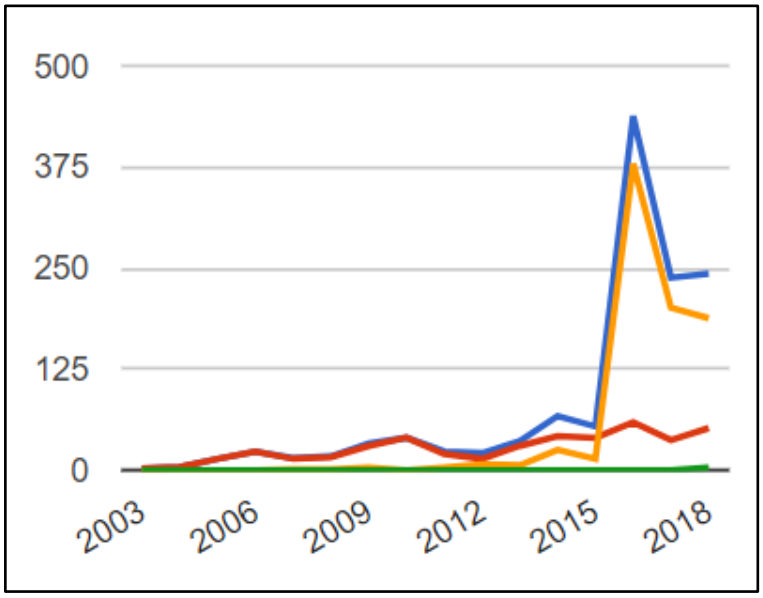


*Figure 5. Dépôts sur HAL (n=1 265, juillet 2018). Bleu: total des dépôts. Jaune: références. Rouge: documents. Verts: annexes.*

Regardons les statistiques d'usage de plus près. Depuis la création de la collection, les 413 documents déposés ont été téléchargés 286 658 fois, avec une médiane de 276 téléchargements. Quant aux 840 notices sans documents, elles ont été consultées 248 401 fois, avec une médiane de 83 (chiffres de mai 2018). Premier constat : tous les documents ont été téléchargés au moins trois fois, et toutes les notices ont été consultées au moins dix fois. On est donc face à un usage du genre "longue traîne", certes avec quelques dépôts très fortement demandés mais surtout avec un grand nombre de documents et de notices occasionnellement consultés et téléchargés. A titre d'exemple, les chiffres de téléchargements pour les documents (figure 6).

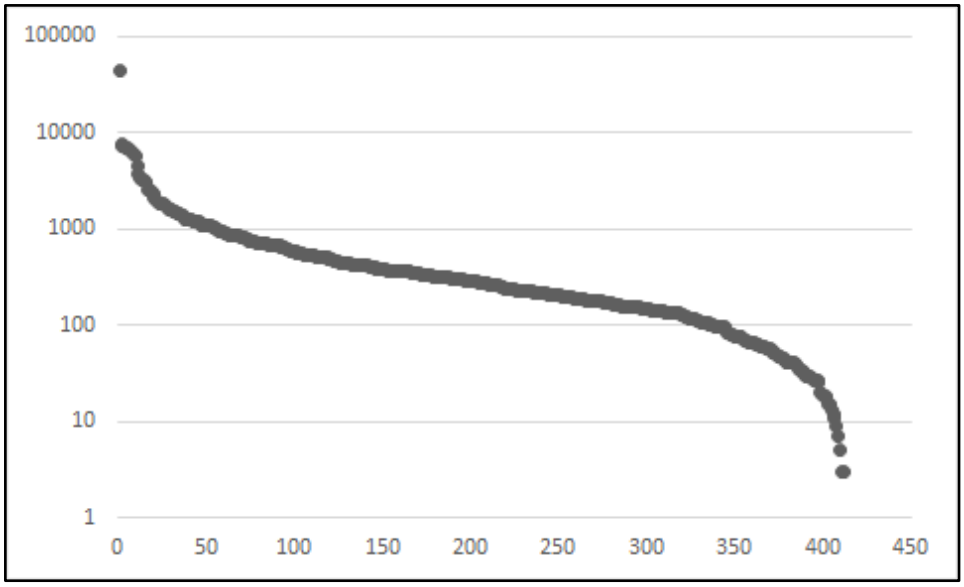


*Figure 6. Téléchargements des documents déposés (n=413, mai 2018, échelle logarithmique)*

A priori, il s'agit d'usage réel, "humain", dans la mesure où l'outil web analytics de HAL filtre l'essentiel de l'activité machine (robots etc.). On peut donc partir du principe que la mise en ligne des documents et notices HAL rend la totalité de la production scientifique du laboratoire visible, et pas seulement une partie.
La date de dépôt joue un rôle, mais l'impact sur l'usage est relatif. A titre d'exemple, les statistiques pour le téléchargement des documents déposés (figure 7).

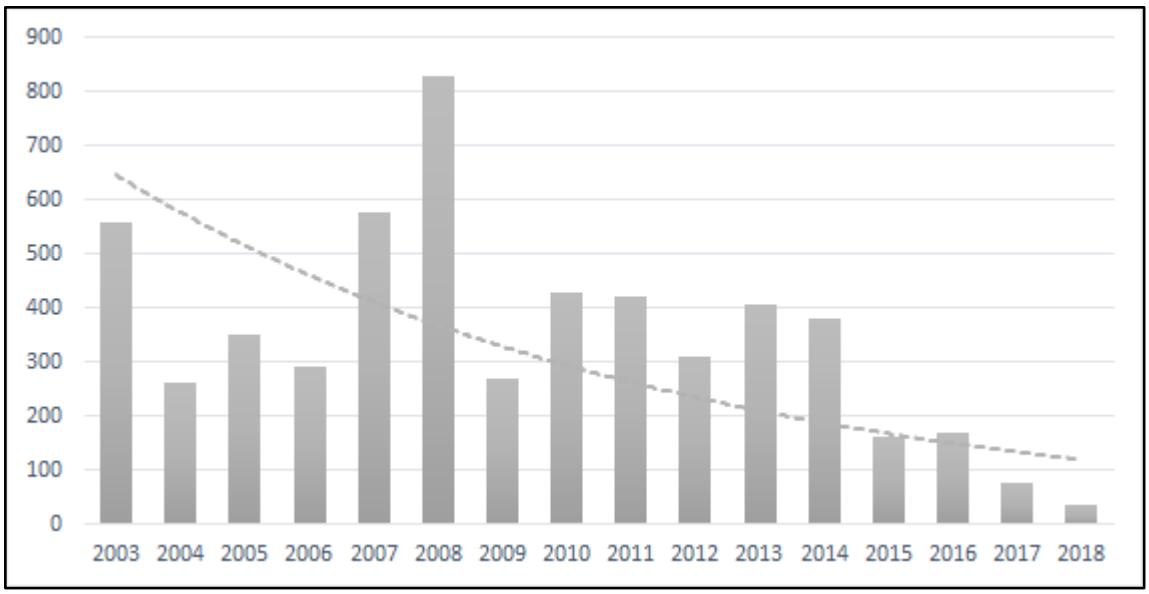


*Figure 7. Le téléchargement des documents en fonction de l'année du dépôt (N=413, mai 2018, médiane)*

Comme la courbe de tendance l'indique, la médiane des téléchargements augmente avec la durée des dépôts. La figure 7 montre aussi que cela s'applique avant tout aux 5-10 dernières années et qu'au-delà, l'impact semble faiblir ; mais ceci est peut-être aussi un effet du petit nombre de dépôts durant les premières années de la collection.
Une dernière observation. Quand on compare les types de documents, on constate quelques différences - là aussi, davantage de tendances que de différences significatives (tableau 2).

| | Nombre | Téléchargements |
|---|---|---|
| Articles | 156 | 273 |
| Communications | 136 | 261 |
| Chapitres | 34 | 283 |
| Thèses | 19 | 288 |
| Rapports | 14 | 300 |

*Tableau 2. Le téléchargement par type de document (mai 2018, médiane)*

On voit ainsi que l'usage moyen des chapitres d'ouvrage, thèses et rapports est en moyenne plus élevé que pour les articles et communications, un phénomène déjà constaté dans le contexte des études sur la littérature grise (Schöpfel & Prost 2016). Mais là encore, il ne faut pas sur-interpréter les chiffres, à cause du petit nombre des rapports, thèses et chapitres.

## Le rayonnement international

Plus haut, nous avons évoqué la part des langues étrangères et notamment de l'anglais qui correspond à un quart des publications. La collection sur HAL fournit deux autres informations sur le rayonnement international du laboratoire.

D'une part, la collection contient un nombre non négligeable de communications dans des conférences nationales et internationales. Ce mode de publication (en conférences) a un impact important sur la dissémination et la valorisation scientifique des travaux de recherche au sein de la communauté scientifique, étant donné le nombre important de chercheurs présents simultanément au moment de la publication. Pour 435 des 450 communications de la collection, la notice contient le pays de la conférence, et il est possible de générer des statistiques d'internationalité (figure 8).

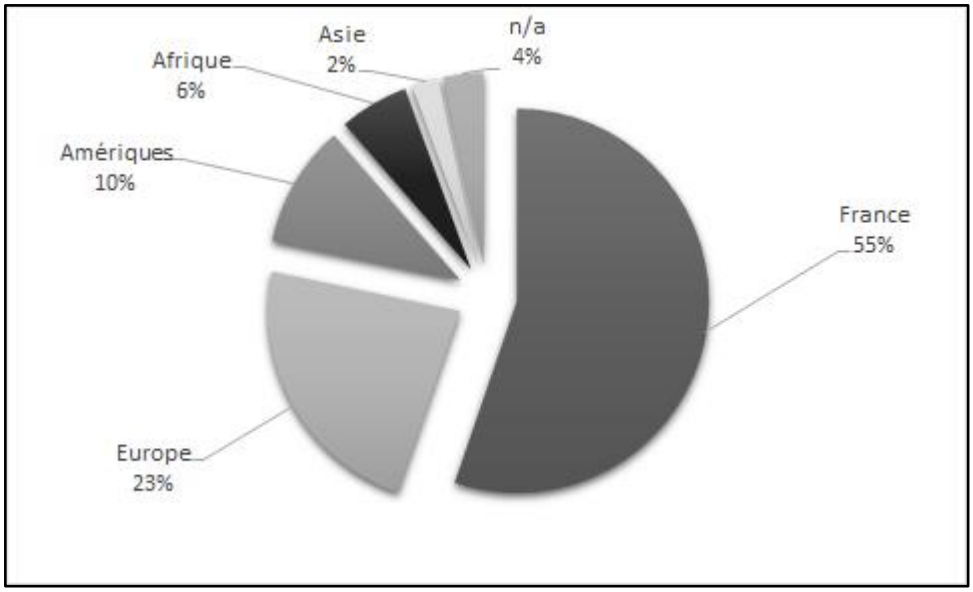


*Figure 8. Pays et continents des conférences de la collection HAL (N=450)*

Plus de la moitié des communications ont été présentées en France ; un quart environ dans d'autres pays européens. Il n'y a pas de tendance claire vers davantage ou moins de communications à l'internationale, et le rapport international/national reste relativement stable dans le temps, en oscillant entre 0,6 et 1,0 par an.
L'information géographique dans les affiliations des auteurs et coauteurs fournit une autre information sur le rayonnement international et les partenariats avec d'autres pays. Vu son intérêt pour l'image du laboratoire, cette information est visualisée sur la page d'accueil de la collection, à l'aide d'une application HAL (figure 9).

*Figure 9. Carte des collaborations, collection GERiiCO sur HAL (N=1 268, 28 juillet 2018)*

La carte est interactive, et on peut modifier les années de publication des références pour visualiser les anciens partenariats ou au contraire, les plus récents, ou pour étudier l'évolution géographique des collaborations.
En revanche, la qualité des résultats de ce "widget" est conditionnée par la qualité des métadonnées des auteurs, en particulier de l'adresse de son affiliation. Nous y reviendrons plus loin.

## Les réseaux des auteurs et coauteurs

Grâce aux informations sur les auteurs et les coauteurs, la collection sur HAL permet aussi une analyse scientométrique des réseaux scientifiques au sein et autour du laboratoire. Ainsi, nous avons mené une analyse des relations entre les différents chercheurs présents dans la collection, à partir des publications communes. Cette analyse a pour objectif de quantifier la communication d'un individu ou d'un groupe, non seulement en termes de volume, mais également de visibilité, d'influence, de partenariats, d'insertion dans les réseaux. Concrètement, chaque chercheur qui est auteur ou coauteur d'au moins un document dans la collection GERiiCO sur HAL est représenté par un nœud dans le réseau. Les relations entre les nœuds représentent une publication commune entre deux chercheurs.
Deux exemples: dans la figure 10, le document est représenté par deux nœuds : [*Joachim Schöpfel, Hélène Prost]* et une relation entre les deux nœuds.


*Figure 10. Exemple d'une publication à 2 auteurs avec 2 affiliations différentes*

Le deuxième exemple (figure 11) est représenté par cinq nœuds *[Romaric Besançon, Stéphane Chaudiron, Djamel Mostefa, Ismaïl Timimi, Khalid Chokri]* et dix relations [*Romaric Besançon__Stéphane Chaudiron, Romaric Besançon__Djamel Mostefa, Romaric Besançon__Ismaïl Timimi, Romaric Besançon__Khalid Chokri, Stéphane Chaudiron__Djamel Mostefa, Stéphane Chaudiron__Ismaïl Timimi, Stéphane Chaudiron__Khalid Chokri, Djamel Mostefa__Ismaïl Timimi, Djamel Mostefa__Khalid Chokri, Ismaïl Timimi__Khalid Chokri*].


*Figure 11. Exemple d'une publication multi-auteur avec trois affiliations différentes*

L’application de cette analyse à l’ensemble des publications de la collection GERiiCO sur HAL aboutit à un réseau avec 592 nœuds et 3 852 relations. Ce réseau global est visualisé par figure 12.

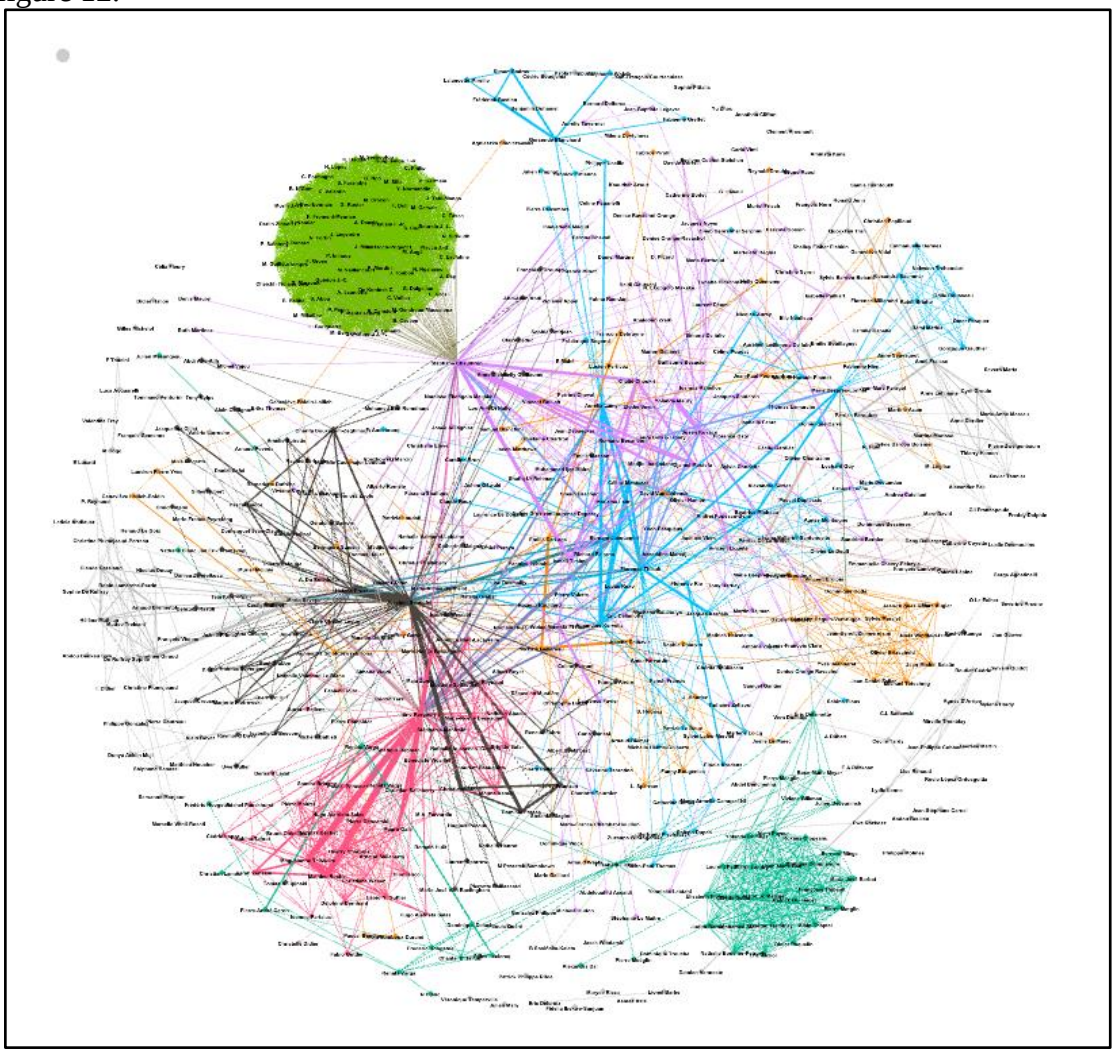

*Figure 12. Réseaux internes et externes entre les chercheurs GERiiCO*

Comme le montre la figure 12, nous remarquons l’existence d’espaces de communautés très variées. Nous divisons cet espace selon deux catégories:

1. *Les communautés fermées* : nous qualifions de *fermées* les communautés ayant un nombre de relations intra-communauté (internes) plus important que le nombre de relations inter-communautés (externes).
2. *Les communautés ouvertes* : nous qualifions de *ouvertes* les communautés ayant un nombre de relations inter-communautés plus important que celui d’intra-communauté.

Nous remarquons aussi que certains nœuds jouent un rôle pivot entre sous-communautés, ce qui permet un meilleur impact et ouverture des travaux de recherche.

Nous n’allons pas plus loin ici dans l’analyse, puisqu’elle sert uniquement d’illustration du potentiel d’une collection de laboratoire sur HAL. Une telle analyse est utile pour la compréhension des liens et dynamiques au sein d’un laboratoire de recherche. Elle est également intéressante pour l’état des lieux dans le cadre du bilan à destination des instances d’évaluation.

# Facteurs de succès et problèmes rencontrés

## Un projet collaboratif

La collection GERiiCO sur HAL a été lancée au printemps 2015, sous une interface customisée et un guichet unique pour l'accès à l'ensemble des dépôts que les membres de GERiiCO avaient effectués antérieurement sur HAL ou l'un des autres portails de HAL (TEL, ArchiveSIC, HAL-SHS, HAL-Lille3). La préparation et mise en œuvre de cette collection a été faite en étroite concertation avec le SCD de Lille 3 et le CCSD, dont les supports de formation à destination des gestionnaires de collection, au moment du déploiement de la version 3 de HAL fin 2014, nous ont été fort utiles. De nouvelles fonctionnalités ont facilité le dépôt de notices : citons par exemple, l'import automatique des principales métadonnées d'un article via le DOI ou la détection immédiate de doublon au moment de la saisie, si la notice existe déjà dans la collection. En 2017, dans le cadre d'une politique de la science ouverte, la Direction Recherche de Lille 3 a recruté une vacataire afin d'accélérer l'alimentation de HAL ; plusieurs jours de son travail ont été consacrés à la collection GERiiCO.

La customisation du site de la collection a été réalisée en partenariat avec la MSH de Dijon qui propose des modèles de mise en page et feuille de style CSS pour les portail et collections sur HAL. Cette coopération avec les instances locales et nationales de l'écosystème HAL a été (et reste) indispensable pour la mise en œuvre et le développement d'une telle initiative.

Les compétences mises en œuvre dans cette phase du projet relèvent du métier de l'information-documentation (conception d'un dispositif documentaire, paramétrage, gestion documentaire, métadonnées etc.). En termes comptables, la mise en place et l'alimentation de la collection n'ont pas eu besoin d'un budget d'investissement ou de fonctionnement particulier. Les seules dépenses sont celles des ressources humaines, en termes de charge de travail, dont une vacation en 2017 (SCD, Direction Recherche, GERiiCO) ; mais il n'y a pas eu d'estimation de coûts, et une partie du travail a été effectuée sur temps personnel.

Après le lancement de la collection, plusieurs membres de GERiiCO, dont la référente publications HAL, ont saisi les références du bilan de 2014 pour obtenir une couverture exhaustive de la production entre 2008 et 2012. La saisie de ces références a pris plusieurs mois et a été terminée fin 2016. Les résultats ont été présentés le 9 février 2017 par Joachim Schöpfel et Hélène Prost dans le séminaire doctoral GERiiCO 2016-2017 "De l'influence en SIC" coordonné par Patrice de la Broise, sous l'aspect "Mesurer l'influence scientifique en termes d'impact : potentiel et limites des nouveaux indicateurs scientométriques". C'était l'occasion notamment de discuter les statistiques d'usage de la collection et de comparer sa visibilité avec Google Scholar et Scopus.

Le débat entre collègues et le soutien de la part du bureau et de la direction du laboratoire sont essentiels pour la réussite d'une telle initiative. Le débat interne, avec la coopération au sein de l'écosystème HAL, est une sorte de garant pour la durabilité de l'initiative à moyen terme.

Les critiques ont porté sur la forme, non pas sur le fond. L'intérêt d'une collection de laboratoire sur HAL n'a pas été remis en question lors des échanges au sein du laboratoire. Les interrogations concernaient d'une part la mise en œuvre (pilotage, gouvernance, interlocuteurs) et d'autre part le résultat, les références mises en ligne (erreurs de saisie, doublons etc.). Dans certains cas, il s'agissait simplement d'une incompréhension de la différence entre dépôt de références (notices) et auto-archivage des publications, et la discussion n'a pas seulement clarifié l'aspect juridique, mais elle a également contribué à une meilleure compréhension des enjeux de la science ouverte.

## Accès vs. signalement

Nous avons indiqué plus haut qu'à l'issue de la saisie des références du bilan 2013-2017, la collection GERiiCO sur HAL comptait 1 253 notices, dont 413 avec le texte intégral (33%). Pour un catalogue de références dont la finalité est de signaler la production d'un laboratoire, c'est beaucoup. Pour une archive ouverte dont la fonction est la communication directe entre chercheurs, ce n'est pas assez.

L'explication de ce taux de 33% est effectivement la saisie exhaustive des références issues des bilans du laboratoire à partir de 2016. Des 350 notices de la collection créées entre 2003 et 2015 par les auteurs, 83% contiennent un lien avec le document déposé sur HAL. Cette situation s'inverse complètement à partir de la saisie des références ; des 903 notices créées entre 2016 et 2018, plus que 22% contiennent ce lien. Pour la plupart, il s'agit de notices créées sous forme d'auto-archivage, par les auteurs. Seulement dans quelques cas isolés, le document a été ajouté à la notice avec l'autorisation des auteurs.

Cet effet d'une stratégie de signalement exhaustif de la production scientifique d'une structure fait débat. Les uns regrettent l'abandon du principe de la communication directe entre chercheurs, d'autres saluent le potentiel pour une évaluation scientométrique en dehors des outils commerciaux (Scopus, Web of Science) ; certains préconisent une distinction nette des dispositifs de l'écosystème de la science ouverte, par une séparation explicite des archives ouvertes (100% accès libre aux documents) et des outils d'évaluation (référentiels).

Pour évaluer notre approche, nous avons comparé la collection GERiiCO avec les autres collections de laboratoire de l'Université de Lille sur HAL.

Pour l'ensemble des 66 laboratoires, le taux de dépôts sur HAL avec documents (texte intégral) est de 20% (19% pour les laboratoires avec une collection HAL, 23% pour les autres), avec une médiane de 32%. La figure 13 montre le rapport entre ce taux et le nombre total des dépôts sur HAL ; la collection de GERiiCO est entourée d'un cercle rouge.

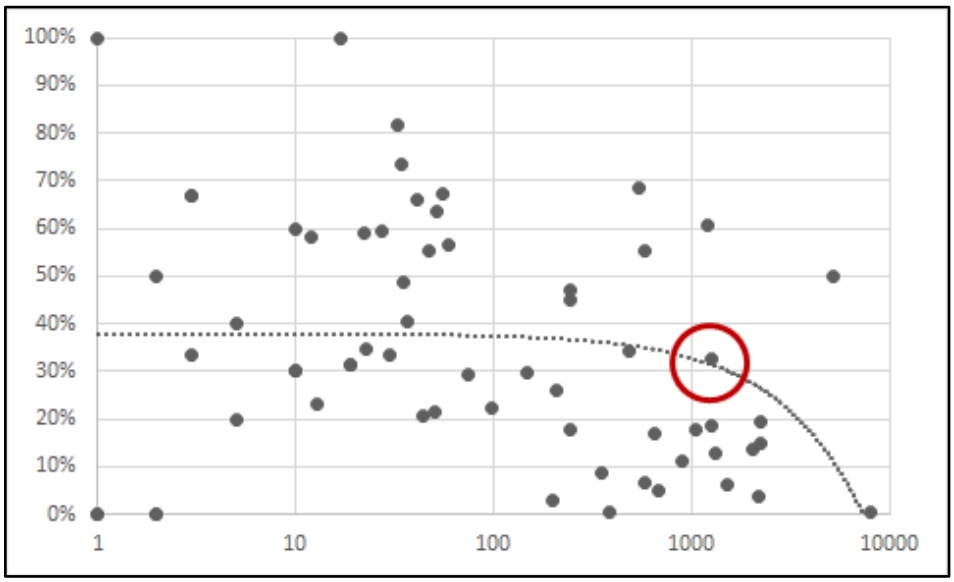


*Figure 13. Taux de dépôts avec document et nombre de dépôts pour les laboratoires de l'Université de Lille (mai 2018, axe du nombre de dépôts en échelle logarithmique)*

Par rapport au nombre de dépôts, la collection de GERiiCO est bien au-dessus de la moyenne des laboratoires de l'Université de Lille (médiane = 52 dépôts). Quant au taux de dépôts avec texte intégral, la collection se trouve juste au-dessus de la moyenne. Par ailleurs, la courbe de tendance visualise un phénomène analysé ailleurs : les grandes collections ont souvent un

taux de texte intégral moins élevé (Prost & Schöpfel 2014), avec des exceptions bien sûr (ici il s'agit de l'UMR 9189 CRIStAL (Centre de Recherche en Informatique, Signal et Automatique de Lille), avec plus de 5000 dépôts et un taux de texte intégral de 50%.
Le même schéma avec uniquement les laboratoires en SHS, tous avec une collection HAL, montre deux groupes avec des approches différentes (figure 14).

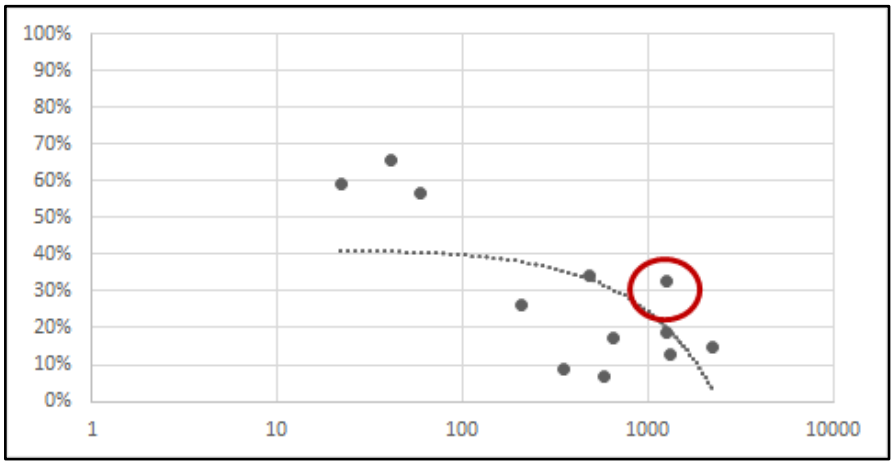


*Figure 14. Taux de dépôts avec document et nombre de dépôts pour les laboratoires SHS de l'Université de Lille (mai 2018, axe du nombre de dépôts en échelle logarithmique)*

La médiane des dépôts est 533, celui du taux de texte intégral de 22%. GERiiCO se trouve donc bien dans le quart supérieur des laboratoires SHS, dans un groupe de neuf laboratoires qui semblent viser comme GERiiCO l'exhaustivité de la collection sur HAL, au détriment du taux de texte intégral. Cependant, un autre groupe de trois laboratoires semble être sur une autre trajectoire, avec un taux élevé de dépôts avec texte intégral mais peu de dépôts. Probablement, ces laboratoires ne font pas (encore) de saisie de références.

## Métadonnées et identifiants

Le problème majeur rencontré et critiqué est la qualité des métadonnées (cf. liste en annexe). Il y a deux problèmes : des métadonnées erronées, et des métadonnées incomplètes ou absentes.

- Métadonnées erronées : souvent des fautes d'orthographe ou d'écriture du nom et prénom d'auteur, de l'affiliation, du titre du de la publication, du nom de la revue, du nom de la conférence, des fautes dans le résumé etc. L'ordre des auteurs est un autre point sensible.
- Métadonnées incomplètes ou absentes : absence d'une affiliation ou affiliation incomplète (absence de pays), oubli d'un auteur, absence de pagination, absence d'un identifiant (DOI ou autre).

D'une manière générale, il peut y avoir deux raisons : soit il s'agit d'une erreur de saisie lors du dépôt et de la création des métadonnées, par l'auteur ou par une autre personne du SCD ou du laboratoire ; soit l'erreur provient de la liste bibliographique du bilan HCERES.
Le contrôle lors de la création des métadonnées est relativement léger, et la procédure de dépôt a été simplifiée pour faciliter et accélérer l'auto-archivage. Il y a donc un risque réel de créer des doublons de notices ou de variantes d'écriture pour les auteurs, les institutions, les revues et conférences.
Lors de la deuxième vague de saisie des références, l'index des noms d'auteurs a fait l'objet d'un contrôle systématique, tout comme les doublons. Aussi, un stagiaire a relevé plusieurs

dizaines d'anomalies de métadonnées dont une partie a déjà été corrigée. De même, le personnel du SCD est intervenu pour encourager les chercheurs à créer leur propre IDHal, l'identifiant auteur unique dans HAL, de façon homogénéiser les métadonnées auteur, tout en permettant au chercheurs de générer leurs curriculum-vitae et bibliographies.

### Indexation

Une autre question se pose pour l'indexation des dépôts. Le principe de l'auto-archivage veut que l'indexation (mots-clés, domaine) soit réalisée par le déposant, c'est-à-dire par l'auteur mais l'ajout de mots-clés n'est pas obligatoire. La plateforme HAL n'a pas de fonctionnalité ou d'outils pour une gestion documentaire de l'indexation. Certes, il est possible de corriger ou compléter l'indexation d'une notice mais il n'y a pas de terminologie contrôlée et/ou structurée, et il n'est pas possible de corriger l'indexation de plusieurs notices à la fois, en remplaçant systématiquement un mot-clé par un autre etc.
Le principe d'une indexation en mode "crowdsourcing" facilite le dépôt et ne devrait pas avoir d'impact sensible sur le référencement par les moteurs de recherche puisqu'ils préfèrent indexer le texte intégral (y compris titre et résumé) et sont capables à faire face à une indexation en langage naturel. En revanche, l'effet n'est pas satisfaisant pour la navigation dans la collection, sur le site de la collection, et pour une analyse scientométrique à partir de l'indexation.
Un problème particulier se pose pour la saisie des références à partir des listes bibliographiques qui ne contiennent ni de mots-clés ni de domaines thématiques. L'option retenue a été une indexation *a minima* du domaine SIC, sans mots-clés, d'autant plus que, depuis le déploiement de la dernière version de HAL, la saisie de mots-clés n'est plus obligatoire pour la validation de la notice.

## Perspectives

La mise en place d'une collection de laboratoire sur HAL contribue à la transformation de l'utilisation individuelle de HAL (l'auto-archivage par les chercheurs) en démarche institutionnelle, avec la création d'un site web, la désignation d'un responsable de collection (référent publications), le suivi des dépôts et la correction des métadonnées. Quelles sont les perspectives d'un tel projet ? Voici quelques éléments de réflexion, par rapport à la nouvelle politique nationale pour la science ouverte, par rapport à la stratégie institutionnelle de l'Université de Lille, et par rapport à l'intérêt d'une telle collection pour l'évaluation des chercheurs et du laboratoire, y compris par la mise en place d'outils à valeur ajoutée.

### Alimentation et mises à jour

Le choix d'une collection HAL vitrine de la production scientifique d'un laboratoire implique une alimentation continue, avec des mises à jour effectuées par les auteurs ou d'autres.
Maintenir le statu quo et continuer comme avant voudrait dire, une alimentation continue par les dépôts des chercheurs avec des saisies de références au moment des bilans, tous les cinq ans. L'inconvénient est que l'auto-archivage par les chercheurs correspond actuellement à environ 30-40% de la production de GERiiCO et que le reste, plus de la moitié, restera donc invisible sur HAL d'un bilan à l'autre. Un autre inconvénient est la charge de travail que cela implique - d'abord la rédaction des listes pour le bilan par les chercheurs, puis la compilation pour le rapport d'activité du laboratoire et en parallèle, la saisie dans HAL. Et comme indiqué, cette procédure est également une source d'erreurs.
On pourrait chercher des alternatives locales, essayer d'inciter les chercheurs à utiliser Zotero pour leur bibliographie ou saisir leur références directement dans HAL, puis les exporter pour

le bilan d'activité. La solution viendra probablement de la nouvelle politique pour une science ouverte, au niveau national aussi bien qu'au niveau institutionnel.

- National : dans le cadre du Plan national pour la science ouverte, il y aura sans doute une incitation forte voire, dans certains cas (financement ANR) une obligation de publier les résultats en open access, dont une partie se retrouvera sur HAL ; en d'autres termes, la part des dépôts devrait augmenter d'ici un ou deux ans. En parallèle, la plateforme HAL devrait se doter de nouvelles fonctionnalités et de nouveaux outils pour faciliter les échanges avec d'autres dispositifs, simplifier l'auto-archivage et rendre l'utilisation de HAL plus attractive.
- Institutionnel : l'Université de Lille a lancé sa propre archive ouverte institutionnelle, LillOA[4], et il n'est pas exclu qu'elle s'aligne sur le modèle de Liège, avec l'obligation d'utiliser ce dispositif pour toute demande d'avancement ou de subvention.

Dans les deux cas, la part de l'auto-archivage devrait augmenter ; si LillOA devient réellement un outil de suivi et d'évaluation de la production des chercheurs de l'Université de Lille, il faudra s'assurer de l'alimentation de la collection HAL à partir de LillOA.
A suivre, donc, et à revoir d'ici un ou deux ans afin d'ajuster la procédure au sein de GERiiCO, le cas échéant. Il faudrait voir également comment récupérer le cas échéant les métadonnées et les documents déposés par un chercheur sur une autre plateforme de son choix. Par un moissonnage systématique ? On ponctuellement, au moment de la saisie des références ? Pour l'instant, il n'y a pas de solution.

## Dépôt du texte intégral

La loi numérique a donné un cadre légal à "l'exploitation secondaire" des publications, y compris de l'auto-archivage sur HAL. Aussi, certains éditeurs autorisent les dépôts des preprints de leurs articles. Les outils de HAL permettent de filtrer ces publications.
Actuellement, la collection GERiiCO sur HAL contient 240 références d'articles sans texte intégral, publiés entre 1993 et 2017. La loi numérique autorise la diffusion en libre accès de ces documents. On pourrait donc imaginer une démarche ciblée pour inciter les chercheurs d'ajouter leur publication à la référence sur HAL, ce qui est une affaire de quelques moments.
Pour 34 références d'articles, dont trois publiés en 2018, l'éditeur autorise l'auto-archivage. De nouveau, on pourrait imaginer une démarche d'incitation pour déposer ces articles - soit par les auteurs, soit par d'autres collègues, avec l'autorisation des auteurs.
De même, on pourrait cibler certains types de documents, dont notamment les rapports (actuellement, 21 notices sans texte intégral) et les communications (actuellement, 307 notices sans texte intégral), et sensibiliser les auteurs de leurs droits et de l'intérêt d'une diffusion des documents en texte intégral, si possible.
Par rapport au dépôt du texte intégral, il faut insister sur la finalité du projet (= vitrine exhaustive de la production scientifique du laboratoire) et sur la différence d'une telle collection avec par exemple une bibliothèque numérique. Ici, il n'y aura pas de filtre ou de contrôle qualité, et la modération de la part du CCSD concerne essentiellement la cohérence des métadonnées, sans évaluer le contenu du document déposé. C'est le principe même du libre accès basé sur l'auto-archivage. En revanche, on pourrait imaginer un dispositif d'annotation et de commentaire a posteriori, une sorte de « post peer review » des documents déposés, à l'instar des épi-revues.

[4] https://lilliad.univ-lille.fr/chercheur/open-access/lilloa-lille-open-archive

## Contrôle des métadonnées

Pour un laboratoire universitaire sans poste de documentaliste, le contrôle systématique des métadonnées de la collection est impossible. Ceci étant, pour améliorer la qualité de la collection et la visualisation des partenariats et collaborations, on pourrait imaginer une sorte d'audit ponctuel pour quelques métadonnées "sensibles" pour identifier et corriger des anomalies et erreurs, en particulier pour les noms d'auteurs, les affiliations et les conférences.
Pour l'instant, le laboratoire invite les chercheurs à vérifier leurs notices et, le cas échéant, à signaler les erreurs à la référente publications, voire les corriger sur HAL.
Le Plan d'action pour la science ouverte annonce plusieurs mesures qui pourraient s'avérer utiles pour améliorer la qualité de ces métadonnées, dont notamment la généralisation d'un identifiant normalisé (ORCID) et l'interconnexion entre HAL et un référentiel national des publications françaises (projet CONDITOR). A voir aussi dans quelle mesure le projet institutionnel autour de LillOA pourra contribuer à la qualité des métadonnées, par exemple par la curation des dépôts.
Faut-il contrôler la qualité des mots-clés, pour obtenir une terminologie homogène et structurée ? Notre réponse est non, pour deux raisons.

- Faisabilité : d'une manière pragmatique, sans documentaliste et sans outils de gestion, vouloir corriger "à la main" la terminologie d'une collection de plus de mille publications est impossible.
- Intérêt : l'intérêt d'une terminologie contrôlée est limitée, aussi bien vis-à-vis du référencement par les moteurs de recherche et autres outils de découverte que vis-à-vis du potentiel d'applications de la fouille de textes et de données (TDM).

Il serait donc plus judicieux d'utiliser un outil TDM pour faire des recherches et pour naviguer dans les données et métadonnées de la plateforme HAL (ou simplement dans la collection GERiiCO), plutôt que d'essayer d'obtenir une qualité de métadonnées sans les moyens et sans véritable intérêt, juste dans la continuité des techniques documentaires des années 80 et 90.
La génération des métadonnées à partir d'un DOI facilite le dépôt et devrait à terme réduire le taux d'erreur, même si cette procédure n'obtient pas non plus une qualité de 100%. De même, on pourrait imaginer un import direct de métadonnées à partir des plateformes d'éditeurs, mais cela impliquerait des partenariats et licences qui n'existent pas au niveau d'un laboratoire.

## Evaluation

La collection HAL d'un laboratoire universitaire est d'un intérêt multiple pour l'évaluation par le HCERES :

- Visibilité : la collection rend la production scientifique du laboratoire visible, bien au-delà de l'indexation par les bases de données scientométrique et d'une façon plus fiable qu'un moteur de recherche.
- Impact : diffuser les publications par la "voie verte", via une archive ouverte, augmente leur impact (consultation en ligne, téléchargement, référencement etc.) et leur procure un "avantage de citation".
- Valorisation : mettre en place un tel outil demande un investissement en temps et énergie, cela prouve que le laboratoire mène une politique réfléchie pour valoriser les résultats de la recherche de ses chercheurs.
- Science ouverte : l'initiative d'une collection sur HAL est le reflet d'une politique locale pour la science ouverte, en conformité avec la politique nationale et la stratégie institutionnelle.

- Scientométrie : la collection permet de faire des analyses scientométriques pointues pour affiner l'auto-évaluation du bilan.
- Suivi et bilan : la collection a le potentiel d'un outil de gestion pour faciliter le suivi et le bilan de la production du laboratoire.

A l'avenir trois initiatives pourraient renforcer l'intérêt pour l'évaluation : rapprocher la typologie documentaire de HAL de la nomenclature du HCERES, améliorer les outils de gestion de HAL pour mieux répondre aux besoins de l'évaluation, et faire le lien de HAL avec les référentiels du gouvernement et avec les systèmes d'information recherche, en particulier avec le futur CapLab. Ces trois initiatives ne relèvent pas du périmètre d'un laboratoire. Ajoutons que la politique nationale actuelle va dans le sens d'une utilisation de HAL pour des fins d'évaluation et devrait favoriser ce genre d'initiatives.

### Lien avec les données de la recherche

A ce jour, la collection ne contient pas de données de la recherche. Il y a quelques liens vers des dépôts de données et un tableau d'enquête en complément du dépôt d'un rapport, mais tout cela reste anecdotique, précurseur peut-être, mais sans réelle importance.

Dans les prochaines années il faut s'attendre à ce que la question des données de la recherche se pose à chaque chercheur, dans le contexte des programmes de recherche européens et français, mais aussi comme bonnes pratiques pour la conservation et la communication des résultats. Le Plan d'action pour la science ouverte de la France prévoit de renforcer les liens entre les publications et les données, par le biais des citations et des liens dans les métadonnées, mais aussi par le biais des revues et articles de données. D'une manière ou d'une autre, la collection des publications de GERiiCO sur HAL sera impactée. Et la question se posera sans doute aussi un jour si GERiiCO se dotera d'une collection de données dans un entrepôt de données, qu'il soit local, national ou international.

## Conclusion

A partir de la collection HAL du laboratoire GERiiCO de l'Université de Lille, nous avons montré comme un laboratoire peut contribuer à la politique de la science ouverte, en mettant en œuvre une stratégie de "voie verte". Nous avons décrit l'intérêt d'une telle approche pour la production scientifique du laboratoire, en termes d'impact et de visibilité, mais aussi pour l'auto-évaluation, pour le bilan d'activité et pour son rayonnement international. Nous avons présenté les facteurs-clés de la réussite d'une telle initiative, dont notamment le soutien par le laboratoire et le partenariat avec le SCD et le CCSD, nous avons évoqué quelques problèmes relatifs aux métadonnées et aux mots-clés, puis nous avons ouvert plusieurs perspectives sur le futur développement d'une telle collection.

Nous n'avons pas mentionné deux points pourtant essentiels pour le positionnement et le développement de HAL - la garantie de conservation à long terme des données (documents) et métadonnées, grâce au partenariat avec le Centre informatique national de l'enseignement supérieur (CINES), et l'interopérabilité de la plateforme avec d'autres dispositifs et infrastructures, grâce à l'application de standards reconnus. Un troisième aspect fait actuellement l'objet d'une évaluation au niveau du Ministère : l'impact d'une telle démarche sur les modèles économiques et perspectives des éditeurs scientifiques en France. Comment concilier l'auto-archivage avec l'avenir de l'édition commerciale ? Peut-on imaginer des partenariats entre HAL et certains éditeurs ?

En guise de conclusion, nous insisterons sur deux autres aspects. Le premier aspect est le potentiel d'une telle collection, dont ses données et métadonnées sont objet et matière primaire de la recherche en sciences de l'information et de la communication. Dans un travail

récent (Mariani et al. 2018), les auteurs ont proposé des mesures permettant d'évaluer l'innovation d'un chercheur dans un domaine scientifique. Pour un auteur ou une publication donnée, il s'agit de repérer les termes novateurs (employés pour la première fois dans un domaine scientifique restreint) et de calculer le nombre d'occurrence de ce terme dans des travaux de recherche ultérieurs dans le même domaine.
D'autres analyses sont possibles, sur les relations et interactions au sein d'une communauté et structure de recherche, sur les partenariats avec d'autres structures et organismes, sur les modes de communication. Les outils de l'analyse des réseaux et de l'exploration des textes et des données ouvrent de nouvelles perspectives à la recherche en SIC; une collection HAL, en tant que corpus enrichi de métadonnées, s'adapte parfaitement à ces outils, à condition d'être exhaustive (ou du moins représentative), d'avoir un nombre significatif de documents en plein texte et des métadonnées de qualité. L'intérêt est double : non seulement alimenter l'auto-évaluation et la discussion stratégique du laboratoire, mais aussi contribuer à la compréhension du travail scientifique et de la production de la connaissance.
Le deuxième aspect est le fait qu'une telle initiative fait partie d'un système (écosystème) plus large. Elle n'a pas de sens comme démarche isolée, mais elle est liée aussi bien à la politique institutionnelle dans les domaines de la science ouverte, du libre accès aux publications et de l'édition scientifique, à la politique nationale de la science ouverte, ses projets, ses financements et priorités, et aux débats et choix au sein des SIC. La conformité du projet local avec les stratégies institutionnelles et politiques lui confère une légitimité indispensable pour un développement dans la durée et permettra aussi, au moins ponctuellement, d'obtenir des moyens supplémentaires pour la saisie et le dépôt, pour la curation et/ou pour certaines analyses.
Faire partie d'un écosystème plus large implique aussi d'anticiper les développements d'autres acteurs et de saisir les opportunités qui se présentent dans l'intérêt du laboratoire, pour augmenter son rayonnement et son impact. Dans ce sens, la contribution au partage des résultats de la recherche devrait figurer parmi les objectifs prioritaires du prochain projet scientifique du laboratoire, tandis que la science ouverte devrait trouver une place privilégiée dans la définition de ses valeurs et de sa vision d'avenir.

## Remerciements

Nous tenons à remercier les collègues et étudiants qui ont contribué à la réussite du projet, en particulier Bernard Jacquemin, Eric Kergosien, Cécile Malleret, Hélène Martin, Clément Plouquet (Université de Lille) et Armelle Thomas (MSH de Dijon).

## Références

Bauin, S., 2014. *L'open access à moyen terme : une feuille de route pour HAL*. CNRS Direction de l'Information Scientifique et Technique, Paris. http://corist-shs.cnrs.fr/Rapport_HAL_DIST_2014
CNRS-DIST, 2016. *Livre blanc : une Science ouverte dans une République numérique*. CNRS Direction de l'Information Scientifique et Technique, Paris. http://books.openedition.org/oep/1548
Dai, Q., Shin, E., Smith, C., 2018. *Open and inclusive collaboration in science: A framework*. OECD science, technology and industry working papers. 2018/07, OECD, Paris. https://doi.org/10.1787/2dbff737-en
Harnad, S., Brody, T., Vallières, F., Carr, L., Hitchcock, S., Gingras, Y., Oppenheim, C., Stamerjohanns, H., Hilf, E. R., 2004. The Access/Impact problem and the green and gold

roads to open access. *Serials Review* 30 (4), 310-314. http://eprints.soton.ac.uk/259939/1/impact.html
Lynch, C. A., 2003. *Institutional repositories: Essential infrastructure for scholarship in the digital age*. 226, ARL Association of Research Libraries. http://www.arl.org/resources/pubs/br/br226/br226ir.shtml
Mariani, J., Paroubek P., Francopoulo, G. (2018). Measuring Innovation in Speech and Language Processing Publications. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan, May 7-12, 2018.
MESRI, 2018. *Plan national pour la science ouverte.* Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation, Paris. http://www.enseignementsup-recherche.gouv.fr/cid132529/le-plan-national-pour-la-science-ouverte-les-resultats-de-la-recherche-scientifique-ouverts-a-tous-sans-entrave-sans-delai-sans-paiement.html
Prost, H., Schöpfel, J., 2014. Degrees of openness. Access restrictions in institutional repositories. *D-Lib Magazine* 20 (7/8). http://www.dlib.org/dlib/july14/prost/07prost.html
Schöpfel, J., Prost, H., 2009. Le JCR facteur d'impact (IF) et le SCImago journal rank indicator (SJR) des revues françaises : une étude comparative. *La Psychologie Française* 54 (4), 287-305. http://dx.doi.org/10.1016/j.psfr.2009.07.002
Schöpfel, J., Prost, H., 2016. Altmetrics and grey literature: Perspectives and challenges. In: *GL18 Eighteenth International Conference on Grey Literature*. November 28-29, 2016, New York Academy of Medicine, New York NY, USA. https://archivesic.ccsd.cnrs.fr/hal-01405443v1
Suber, P., 2012. *Open access*. MIT Press, Cambridge Mass. http://mitpress.mit.edu/books/open-access
Thirion, P., Rentier, B., 2014. La communication scientifique à la croisée des chemins : Enjeux et stratégies institutionnels. In: *Open Access Week*, 22 octobre 2014, Université de Strasbourg. https://orbi.uliege.be/handle/2268/173664

# Annexe

## Métadonnées sur HAL

Chaque document est décrit par un ensemble de 31 variables :

- *halId_s* : l'identifiant du document sur la plateforme HAL.
- *contributorid_i, contributorFullName_s* : l'identifiant et le nom de la personne qui a fait le dépôt sur HAL.
- *Version_I* : la version du document
- *Uri_S* : l'URL du document
- *docType_s* : cette variable peut prendre une valeur parmi les 13 valeurs suivantes: COMM, ART, COUV, DOUV, OTHER, REPORT, UNDEFINED, THESE, POSTER, OUV, MEM, HDR, PATENT.
- *doiId_s, nntId_s* : l'identifiant doi
- *Title_S, subTitle_s* : le titre du document et le sous-titre s'il existe
- *authId_i, authFullName_s, authLastName_s* : les iidentifiants, les noms et les prénoms de l'ensemble des auteurs.
- *producedDate_s* : la date de publication
- *Domain_S*: le domaine qui peut prendre une valeur parmi les 4 valeurs suivantes : 0.shs (Sciences Humaines et Sociales), 1.shs.info (Sciences de l'Information), 2.shs.info.doc (Sciences de l'Information et de la documentation) ou 2.shs.info.com (Sciences de l'information et de la Communication).

- *journalTitle_s, journalPublisher_s, Volume_s, Number_s, Page_s* : lorsqu'il s'agit d'un article de revue ces champs contiennent respectivement, le nom du journal, le nom de l'éditeur, le numéro et le nombre du volume, les numéros de pages.
- *conference_Title_s, conferenceStartDate_s, Country_s* : pour les communucations dans des congrès ces champs contiennent le nom de la conférence, la date de début et le pays.
- *Language_s* : la langue du document
- *inPress_book:* 2 valeurs possibles : *oui* si le document est apparu dans un livre de presse sinon *non*.
- *fileType_s :* le type du fichier
- *Nb_consultation_fiche, Nb_consultation_doc :* le nombre de consultations de la fiche et du document
- *submittedDate_s, submittedDateY_i* : la date du dépôt du document sur HAL.