Analyse et situation des sciences de l'information et de la communication dans les pratiques d'archivage

Analysis of information and communication sciences in archiving practices

Alaric TABARIES (1) David REYMOND (2)

(1) IMSIC, Université de Toulon
alaric-tabaries@etud.univ-tln.fr
(2) IMSIC, Université de Toulon
dreymond@univ-tln.fr

Résumé. Le mouvement de la science ouverte se pose comme le nouveau paradigme de référence de diffusion du savoir au sein de la communauté scientifique. En France, l'archive électronique HAL occupe une place centrale pour l'archivage et la diffusion des résultats de recherche. À ce jour, plus de trois millions de notices bibliographiques sont accessibles directement ou par API - dont plus d'un million proposent un texte intégral. L'analyse des captas via l'API nous permet de décrire les pratiques d'archivage, questionnant alors les fonctions élémentaires de l'archive HAL. Au-delà des différences disciplinaires constatées, et plus particulièrement de la singularité des sciences de l'information et de la communication dans ce mouvement, nous montrons que la qualité du processus de dépôt tend à être négligée alors que ce dernier influe directement sur la diffusion du savoir. Nos résultats montrent toutefois que l'accompagnement du chercheur vers une meilleure compréhension des enjeux de la science ouverte permet d'apporter une réponse à cette problématique.
Mots-clés. Science ouverte ; HAL ; pratiques d'archivage ; auto-archivage ; sciences de l'information et de la communication

Abstract. The open science movement has emerged as the new paradigm for disseminating knowledge within the scientific community. In France, the HAL electronic archive plays a central role in the archiving and dissemination of research results. To date, more than three million bibliographic records are listed - over a million of which are full-text. The analysis of this information allows us to describe archiving practices, and to question the basic functions of the HAL archive. Beyond the disciplinary differences observed, and more particularly the singularity of the information and communication sciences in this movement, we show that the deposit process tends to be neglected, even though it has an influence on the very process of knowledge dissemination. Our results show, however, that supporting researchers towards a better understanding of the challenges of open science can provide a solution to this problem.
Keywords. Open science ; HAL ; archiving practices ; self-archiving ; information and communication sciences

1 Introduction

Le mouvement de la science ouverte se pose comme le nouveau paradigme de référence de diffusion du savoir au sein de la communauté scientifique. En France, l'archive électronique HAL occupe une place centrale pour l'archivage et la diffusion des résultats de recherche (Berthaud et al., 2021), les notices que l'archive héberge étant moissonnées par de nombreux moteurs de recherche scientifiques (*Visibilité des dépôts HAL*, 2023). À ce jour, plus de trois millions de notices bibliographiques sont listées - dont plus d'un million associées à un texte intégral - et le rythme de dépôt de nouvelles ressources ne cesse de croître. À l'heure des sciences sociales de « troisième génération » (Boullier, 2017), l'étude quantitative de ces notices revêt un intérêt particulier dans l'analyse de l'appropriation de cet outil de médiation documentaire, inférée sur les activités des chercheurs, et, par extension, de l'impact du processus de dépôt sur la diffusion du savoir scientifique.
Nous décrivons ici, dans un premier temps, le contexte scientifique de l'analyse de l'archivage de la production scientifique sur l'archive institutionnelle. À l'aide d'une collecte exhaustive, nous produisons ensuite un panorama national montrant l'étendue des pratiques d'archivage, leurs variétés et leurs dynamiques. Nous réalisons alors une étude comparée des sciences de l'information et de la communication (SIC) avec les sciences humaines et sociales (SHS). Enfin, nous étudions, puis comparons l'impact des pratiques d'archivage sur le processus même de diffusion scientifique. Nos conclusions montrent une prédominance historique des SIC dans la pratique de l'archivage documentaire, qui s'explique par l'archive éponyme, mais qui se dilue au fil des temps pour se perdre au niveau des SHS.

2 Contexte

Initiée en 2001 par le CNRS dans un contexte de fondement du mouvement de la science ouverte (Berthaud et al., 2021), l'archive ouverte HAL, mise en avant par les plans nationaux en faveur de la science ouverte (MESRI, 2019, 2021), occupe aujourd'hui une place centrale pour l'archivage et la diffusion des résultats de recherche en France (Berthaud et al., 2021). Singularité en SHS, le projet @rchiveSIC a été établi en mai 2002 en réponse au colloque intitulé "Place et enjeux des revues pour la recherche en Infocom (SFSIC)", tenu le 25 mars 2002. Ces initiatives s'inscrivent dès lors dans le contexte du mouvement mondial des archives ouvertes, dont les prémices ont été établies par Paul Ginsparg au début des années 90 au *Los Alamos National Laboratory* par la création du site Arxiv. L'initiative de Paul Ginsparg a eu un impact significatif sur le paysage de l'édition scientifique et technique en introduisant la possibilité d'un modèle alternatif plus diversifié. Ces vingt dernières années, les politiques locales et nationales ainsi que les appels en faveur de ce mouvement se multiplient (*Dates clés de la science ouverte*, 2023) et posent désormais la science ouverte comme paradigme de référence de diffusion du savoir. Toutefois la science ouverte se développe de façon inégale entre les sciences et des différences disciplinaires sont constatées (Larrieu & Schöpfel, 2022) singularisant de fait la question de l'accompagnement au processus d'auto-archivage, progressivement inclus dans la mission de *diffusion* de l'enseignant chercheur. Ainsi, par l'utilisation plus forte de livres (chapitre, ouvrages, direction) et le français comme première langue, les SHS présentent les taux d'accès ouverts les plus faibles. À travers ArchiveSic, au sein des SHS, les SIC disposaient dès la création, selon les cofondateurs (Chartron, Noyer, Gallezot), "d'une piste favorisant le développement scientifique de notre communauté" pour rendre visibles et accessibles les travaux. Comment s'est

finalement développée cette piste et quelles sont depuis les dynamiques de notre communauté en termes d'auto-archivage ? La documentation comme sous-discipline des SIC fait-elle figure singulière au sein des SHS ? Pour mener cette étude, nous analysons l'ensemble de l'archive HAL depuis sa création. Nous procédons à une première caractérisation qualitative pour déterminer à l'aide de travaux fondateurs les types et logiques de dépôts. Les notices récupérées sont enrichies d'indicateurs bibliométriques pour produire des axes d'analyses que nous présentons dans la section résultats.

3 Qui dépose, selon quel objectif et à quelle fréquence ?

Des travaux qualitatifs ont été menés pour étudier la contribution d'une partie de l'archive HAL. Mahé & Prime-Claverie (2017a, 2017b) ont caractérisé les dépôts HAL en SHS selon les notions de niveaux de contributions, logiques de dépôt, et, habitude de dépôt.

3.1 Contributeurs

Les autrices ont ainsi identifié en premier lieu quatre types de contributeurs qui permettent de distinguer l'entité à l'origine de l'acte de dépôt et distinguer ainsi des niveaux de contribution. Elles distinguent ainsi (Mahé & Prime-Claverie, 2017a) l'auto-archivage (l'auteur dépose ses travaux) ; l'intermédiaire chercheur mais non auteur du dépôt ; l'intermédiaire non chercheur et le dépôt automatisé (import de bases de données le plus souvent).

Plus récemment, Schöpfel et al. (2023) ont étudié les dépôts sur l'archive HAL de 1 246 laboratoires affiliés aux 10 plus grandes universités françaises. Ils ont réalisé une typologie détaillant sept niveaux de contributions différents. Nous verrons par la suite, qu'au travers de cette étude, nous ne sommes pas en mesure d'apporter une caractérisation aussi fine au volume de données analysé.

3.2 Logiques et habitudes de dépôt

La notion de logique de dépôt (tableau 1) est une notion qui dépend de l'écart entre la date de publication et la date de dépôt ainsi que de l'association (ou non) du texte intégral à une notice. Ainsi, lorsqu'un dépôt est effectué dans l'année suivant la publication du document, on parle de logique de communication scientifique directe (avec texte intégral) ou de logique de référencement (sans texte intégral). À l'inverse, on parle de logique d'archivage (avec texte intégral) et de logique de recensement (avec texte intégral) lorsque l'écart de temps est plus conséquent. Le tableau 1 reprend ces logiques de dépôt, selon Mahé & Prime-Claverie (2017a).

La notion d'habitude de dépôt indique si le contributeur est nouveau, régulier ou encore sortant (2 ans sans dépôt) en s'appuyant sur sa fréquence de dépôt.

	Dépôt récent (moins d'un an)	**Dépôt ancien**
Dépôt avec fichier	Communication scientifique immédiate	Archivage
Dépôt sans fichier	Visibilité de front de recherche (référencement)	Visibilité des recherches antérieures (recensement)

Tableau 1. *Logiques de dépôts (Mahé & Prime-Claverie, 2017a)*

3.3 Complétion des dépôts

L'indexation est le processus de description et de représentation d'une ressource documentaire afin de permettre sa recherche et son repérage ultérieur. Au cours de

ce processus, les métadonnées complètent le document indexé et enrichissent par un niveau supplémentaire le document connecté (Broudoux, 2015). Les métadonnées fournissent les éléments clés pour indexer de manière cohérente les ressources, en identifiant et en décrivant les caractéristiques essentielles de chaque document devenant ainsi indispensables pour l'indexation, la diffusion et le partage des connaissances. En guidant l'indexation, elles facilitent la recherche et la découverte de ressources, améliorent l'efficacité de la gestion de l'information, favorisent l'interopérabilité entre les systèmes et les sources de connaissances, et contribuent à la préservation et à la pérennité des connaissances à long terme (Alhuay-Quispe et al., 2017 ; Beall, 2005 ; Jaffe, 2020 ; ou-Pair, 2005 ; Park, 2009).

Ainsi, nous avons pour notre part, proposé une métrique permettant d'évaluer le niveau de description d'une notice en se basant sur la complétude des métadonnées renseignées, rendant compte d'une vue synthétique de la description des notices (Tabariès, 2022). Toutefois, nous verrons qu'à ce stade cette métrique n'est pas encore utilisable, la question n'est pas encore de l'accompagnement qualitatif du chercheur à l'auto-archivage mais simplement pour le moment de son accompagnement différencié sur le plan disciplinaire.

3.4 Accès libre et métriques d'impact

Un nombre conséquent de travaux questionnant une possible corrélation entre accès ouvert et métriques d'impact ont été conduits (Swan, 2010). Les résultats obtenus varient, une fois encore, en fonction des disciplines étudiées mais les corrélations constatées semblent s'effacer progressivement de par l'utilisation croissante du site web Sci-Hub (Maddi & Sapinho, 2022). Une alternative non légale aux circuits standards de l'édition qui trouverait une issue dans l'auto-archivage généralisé.

L'ensemble de ces travaux soulignent des spécificités dans le processus de dépôt au niveau disciplinaire qui sous-tendent de nécessaires déclinaisons des processus d'accompagnement.

4 Méthodes

4.1 Périmètre de recherche

En premier lieu, nous interrogeons l'interface de programmation mise à disposition par le portail HAL. Nous récupérons alors, progressivement, les métadonnées associées décrivant les notices référencées sur l'archive, pour les téléverser vers une base de données locale. Le périmètre temporel de cette étude s'étend à l'intégralité des notices déposées entre le premier janvier 2002 au 31 décembre 2022 inclus. Le corpus d'étude contient les métadonnées de 3 158 018 notices, récoltées à la date du 5 janvier 2023. Nous utilisons les métadonnées pour effectuer des traitements basiques. Ainsi, pour chaque notice, nous vérifions la présence de champs descriptifs (fichier, mot-clé, résumé) indépendamment du langage ; nous attribuons un score mesurant le niveau de description de la ressource (Tabariès, 2022). Ces métadonnées sont enrichies par les métriques d'usage et d'impact internes à HAL, c'est-à-dire, les nombres de visionnages et de téléchargements, en moissonnant directement la page Web de la notice et, enfin, lorsque l'identifiant DOI est associé à une notice, nous interrogeons l'API Dimensions pour récupérer des métriques d'usage classiques en bibliométrie : le nombre de citations, le *relative_citation_ratio* (RCR) et le *field citation ratio* (FCR). Le FCR est la moyenne normalisée du taux de citation par revue dans le même domaine. Ainsi, il permet de mesurer l'impact d'un article spécifiquement dans son domaine

scientifique. Le RCR est calculé en comparant le taux de citations d'un article au taux de citations attendu (le FCR) dans le même domaine de recherche et publiés au cours de la même période. Il permet donc de mettre en perspective l'impact d'un article indépendamment du temps (Janssens et al., 2017). Ces dernières sont disponibles pour seulement 807 145 notices. Les métriques d'impact sont récoltées entre le 7 janvier et le 11 janvier 2023.

Nous retirons les notices déposées via des imports automatisés pour cibler les dépôts caractérisant les pratiques humaines de l'archivage. Nous créons enfin deux sous corpus d'étude d'un côté les notices relevant des SHS et, de l'autre, celui spécifique aux SIC. Ce filtrage s'effectue à l'aide du champ spécifique des métadonnées (*domain_s*). Le premier corpus comprend 897 970 notices dont 56 892 (environ 6%) listant des métriques d'impact, le second comprend 33 723 notices dont 1 758 (environ 5%) listant des métriques d'impact.

4.2 Caractérisation des dépôts

Nous caractérisons les notices selon les notions de logiques de dépôt et de niveaux de contributions tels que définies par Mahé & Prime-Claverie (2017a). Ainsi, pour la qualification des notices en termes de logiques de dépôt, nous suivons la méthodologie décrite par les auteurs. Nous mesurons l'écart de temps entre la date de dépôt de la notice et la date de publication du document scientifique associé. Lorsque le dépôt est effectué dans l'année suivant la date de publication et qu'un document est joint, on estime que ce dernier est réalisé dans une logique de communication scientifique directe. Si toutefois le document n'est pas joint, on estime que l'acte est réalisé dans une logique de référencement. À l'inverse, pour un dépôt effectué plus d'un an après la date publication du document, on parle d'archivage quand un document est présent et de recensement quand cette information s'avère manquante. Nous appliquons cette méthode sur les notices déposées jusqu'à l'année 2021 pour ne pas induire un biais dans les résultats.

Pour la notion de niveau de contribution, la volumétrie d'informations à traiter étant conséquente, nous nous sommes écartés du protocole proposé par les auteurs, dans le but d'en automatiser le traitement. Nous nous fondons sur l'information communiquée par l'interface de programmation HAL indiquant si la notice a été déposée en auto-archivage ou non (champ *selfArchiving_bool*), c'est-à-dire par un des auteurs. Lorsqu'il ne s'agit pas d'auto-archivage, nous comparons le nom de l'entité déposante à une base de données de noms d'individus dérivée de réseaux sociaux (Remy, 2021) pour différencier le dépôt humain d'un dépôt automatisé, utilisé lors de l'import de données en nombre important. Toutefois, nous ne sommes pas en mesure d'affiner la notion de dépôt effectué par un intermédiaire (intermédiaire chercheur et non chercheur).

Enfin, nous caractérisons les contributeurs déposant les notices toujours selon la notion développée. Lorsqu'un contributeur dépose pour la première fois et dépose également dans les deux années qui suivent le dépôt initial, il est alors qualifié comme nouveau contributeur pour plusieurs années. A l'inverse, un contributeur qui ne dépose aucune autre notice dans la période de deux ans suivant le dépôt initial est considéré comme un nouveau contributeur pour une année seulement. Les contributeurs réguliers qui ne publient plus pendant deux années sont considérés comme sortants.

4.3 Analyse des relations entre pratiques d'archivage et métriques d'impact (processus de diffusion)

Nous sélectionnons, dans un premier temps, uniquement les notices déposées entre 2014 et 2018 dans le but de ne pas intégrer de documents dont les métriques

d'impact pourraient évoluer fortement. Nous constituons ensuite des échantillons aléatoires composés de 4 000 notices pour chaque groupe étudié. Nous appliquons un test statistique T de Welch sur les groupes afin de vérifier les hypothèses formulées. À noter que nous utilisons le FCR étant donné que l'information est disponible pour un nombre plus important de notices que le RCR.

5 Résultats

Les deux corpus ainsi qualifiés nous permettront de comparer les distributions historiques de dépôt selon les différentes modalités qualitatives précédentes.

5.1 Panorama des pratiques d'archivage en sciences humaines et sociales

Le corpus de notices relevant des SHS comporte majoritairement des articles (~36%), des communications (~21%) et des chapitres d'ouvrage (~21%). Ces dernières années, nous constatons une accélération des dépôts d'articles au détriment des communications.

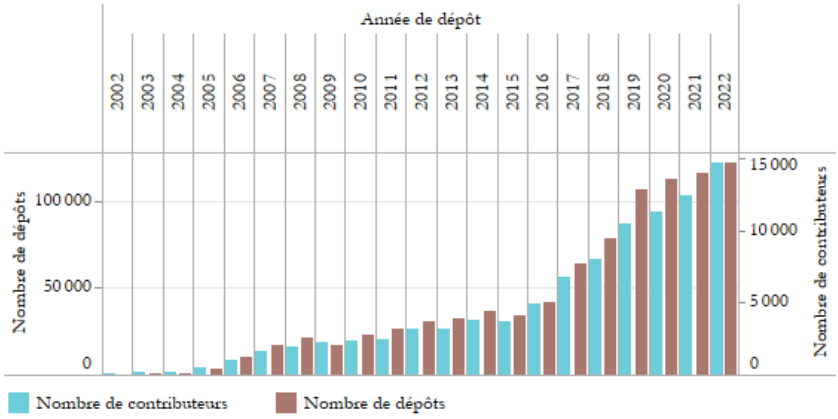

Figure 1. *Évolution du nombre de dépôts et du nombre de contributeurs relevant des SHS*

Les nombres de notices déposées et de contributeurs uniques (fig. 1) ne cessent de croître depuis la création de l'archive. Nous distinguons deux périodes distinctes : jusqu'en 2015, le nombre de contributeurs uniques croît lentement alors qu'à partir de 2016, ce nombre augmente rapidement. Le nombre de dépôts annuels ne suit pas tout à fait la même évolution, cela est probablement dû à un rattrapage des documents déjà produits mais non déposés jusqu'alors. Ce changement coïncide avec la promulgation de la Loi pour une République numérique (2016).

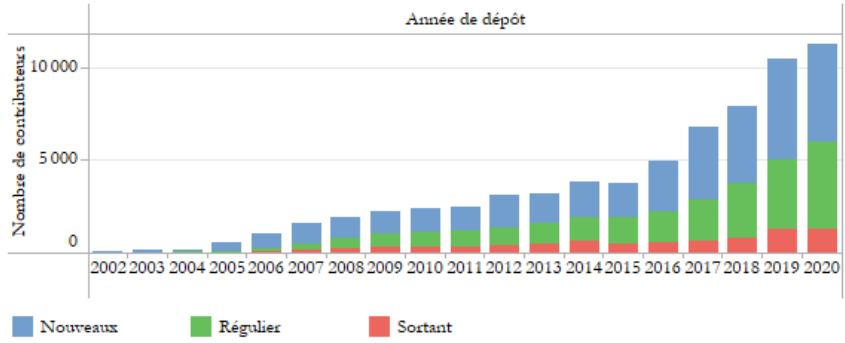

Figure 2. *Évolution des flux de contributeurs relevant des SHS*

Sans surprise, les nombres de nouveaux contributeurs et de contributeurs réguliers (fig. 2) suivent la même évolution que celle décrite précédemment. Nous observons trois phases, la première pour laquelle l'utilisation de la plateforme d'archivage progresse lentement depuis la création avec moins de 1000 contributeurs « précurseurs » avant 2006. Arrive alors une seconde phase d'adoption par l'arrivée lente mais régulière d'usagers pour pratiquement atteindre les 4000 contributeurs en 2016. La Loi pour la République numérique marque le début d'une phase exponentielle de contributions pour en dénombrer plus de 11000 en SHS en 2020. Nous notons que la grande majorité des contributeurs déposent en auto-archivage (~90%), les contributeurs intermédiaires ne représentent donc qu'environ 10%. Ces répartitions restent stables dans le temps. Nous observons qu'en 2022, un contributeur intermédiaire dépose environ 22 notices contre 6 pour les contributeurs en auto-archivage.

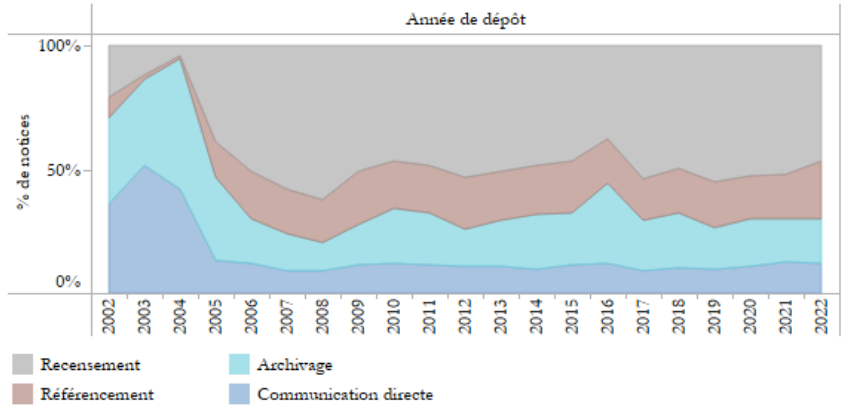

Figure 3. *Évolution des logiques de dépôt pour les notices relevant des SHS*

Les notices déposées selon une logique de recensement (fig. 3) représentent la majorité des dépôts (~47% en 2022), viennent ensuite celles déposées selon une logique de référencement (~23% en 2022), puis, celles déposées selon une logique d'archivage (~18% en 2022) et enfin, celles déposées selon une logique de communication scientifique directe (~12% en 2022). Ces répartitions restent assez

stables ces dernières années à l'exception d'une augmentation soudaine pour les dépôts effectués dans une logique d'archivage avec environ 32% en 2016.

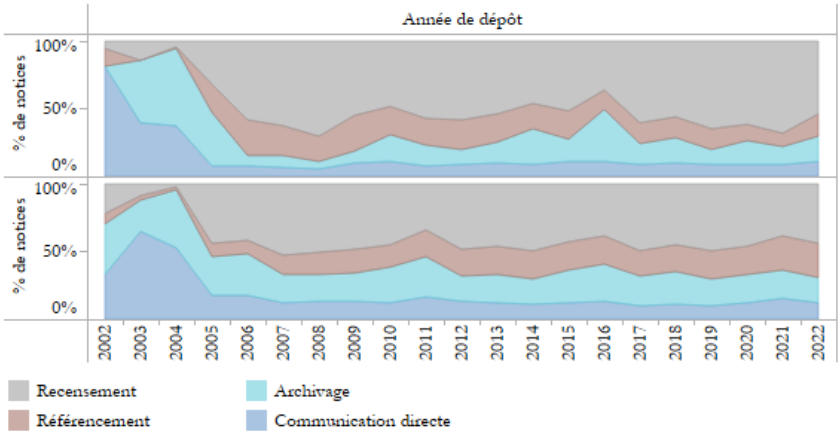

Figure 4. *Évolution des logiques de dépôt pour les notices relevant des SHS déposées par des intermédiaires (haut) et en auto-archivage (bas)*

Les notices déposées en auto-archivage (fig. 4) suivent plus des logiques de référencement que les dépôts effectués par des intermédiaires. Les contributeurs intermédiaires s'attachent donc plus à recenser les travaux antérieurs, sans dépôt de texte intégral.

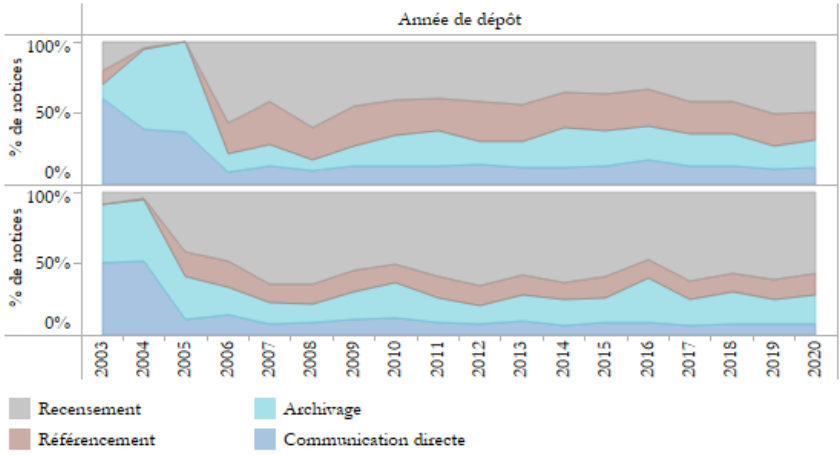

Figure 5. *Évolution des logiques de dépôt pour les notices relevant des SHS pour les contributeurs réguliers (haut) et nouveaux (bas)*

Nous observons également que les contributeurs réguliers soumettent plus selon des logiques de communication scientifique directe (~12% contre ~8% pour les nouveaux contributeurs en 2020).

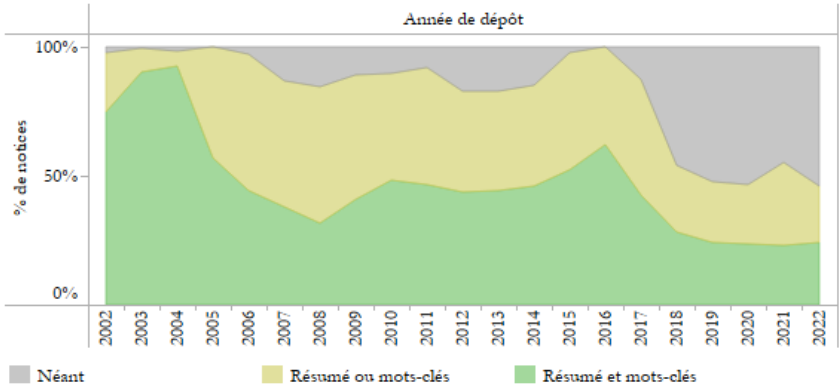

Figure 6. *Évolution du niveau de description pour les notices relevant des SHS*

Nous distinguons une nette baisse dans la qualité de description des notices déposées (fig. 6), selon une notion de complétude des deux métadonnées clés (résumé et mots-clés), dès 2016. Cette baisse est corrélée avec la forte augmentation du nombre de dépôts annuel. Ainsi, en 2022, la majorité des notices (~54%) ne présentent ni résumé ni mot-clé. Le lecteur intéressé trouvera une analyse approfondie dans nos précédents travaux (Tabariès, 2022).

Voyons maintenant si la logique de dépôt a une incidence sur la qualité ou le soin apporté à ces dépôts. La figure 7 montre l'évolution de ce degré description pour les notices relevant des SHS selon des logiques de communication scientifique directe et d'archivage en haut et celles suivant des logiques de référence et de référencement en bas)

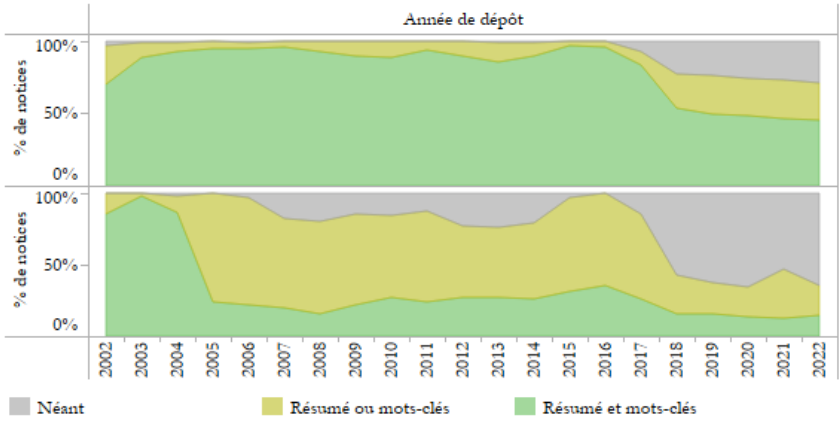

Figure 7. *Évolution du niveau de description pour les notices SHS selon des logiques de communication scientifique directe et d'archivage (haut) et suivant des logiques de recensement et de référencement (bas)*

Les dépôts effectués selon des logiques de communication scientifique directe ou d'archivage sont mieux décrits que les dépôts effectués selon des logiques de référencement et de recensement : en 2022, environ 45% des dépôts qui suivent des logiques de communication scientifique directe et d'archivage présentent à la fois un résumé et des mots-clés alors que cette proportion est d'environ 15% pour ceux effectués dans des logiques de référencement et de recensement. Cet état de fait

conforte plusieurs éléments : d'une part la logique d'archivage et de communication est soucieuse de la qualité de l'indexation, nous parlons ici de lisibilité, en opposition avec les deux autres logiques de référencement et de recensement qui marquent une simple préoccupation de visibilité (Reymond, 2022).

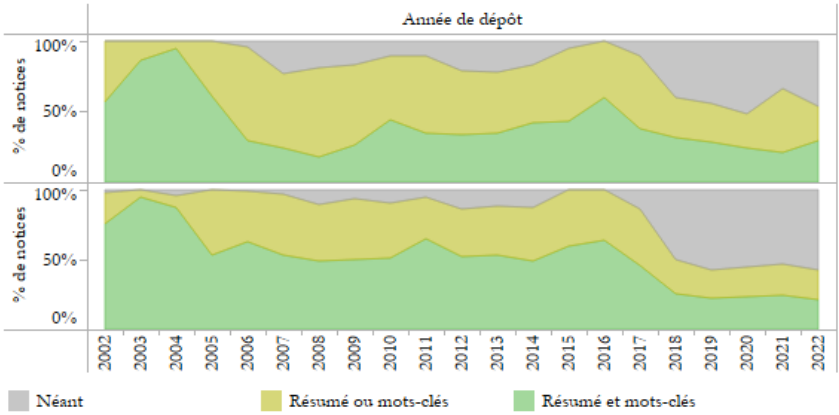

Figure 8. *Évolution du niveau de description pour les notices SHS déposées par des intermédiaires (haut) et en auto-archivage (bas)*

Nous constatons que la tendance s'est inversée avec le pic de 2016 (fig. 8) : les notices déposées en auto-archivage sont mieux décrites que celles déposées par des intermédiaires ces dernières années. Ainsi, en 2022, environ 47% des notices déposées en auto-archivage ne présentent ni résumé ni mots-clés contre, environ, 57% des notices déposées par un intermédiaire. L'intermédiaire, disposant plus difficilement ces informations, dépose dans une logique de visibilité (ou pour répondre aux recensements métriques des évaluations de type HCERES).

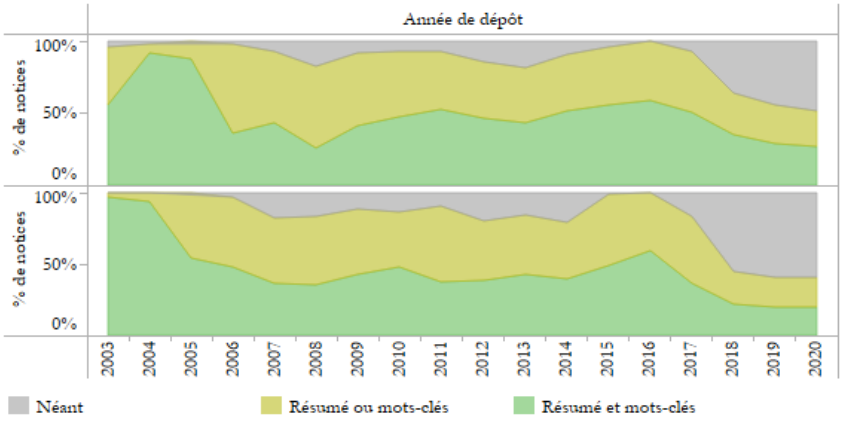

Figure 9. *Évolution du niveau de description pour les notices relevant des SHS déposées par des contributeurs réguliers (haut) et par de nouveaux contributeurs (bas)*

Pour conclure, nous observons que les contributeurs réguliers décrivent mieux les notices lors du processus de dépôt (fig. 9). En effet, en 2020, environ 27% des notices déposées par des contributeurs réguliers sont décrites à la fois par un résumé

et par des mots-clés, contre, environ 19 % par de nouveaux contributeurs. Au cours des trois phases nous constatons des variations notables dans les modalités de contribution. La première phase (jusqu'en 2006) est marquée par l'arrivée des précurseurs intermédiaires ou en auto-archivage ciblant communication et la pérennisation de leurs travaux en apportant une dimension qualitative notable : résumés et mots sont le plus souvent présents. La logique de dépôt prend un tournant dès 2006 et devient essentiellement du recensement (50%) pour tout type de contributeurs, maintenant la présence de résumé ou de mots clés seulement. La troisième phase est marquée par une diminution drastique de la qualité de l'indexation par une quantité de contributions (plus de 60%) ne présentant aucune de ces deux métadonnées.

Voyons maintenant s'il est des spécificités propres aux SIC.

5.2 Les sciences de l'information et de la communication au sein des SHS

Les SIC sont une branche des SHS qui, depuis Jean Meyriat, inclue les sciences de la documentation (*Library and Information Science*) et dénotent une singularité de la pratique de l'archivage et du référencement. Nous montrons ci-après les disparités les plus marquées.

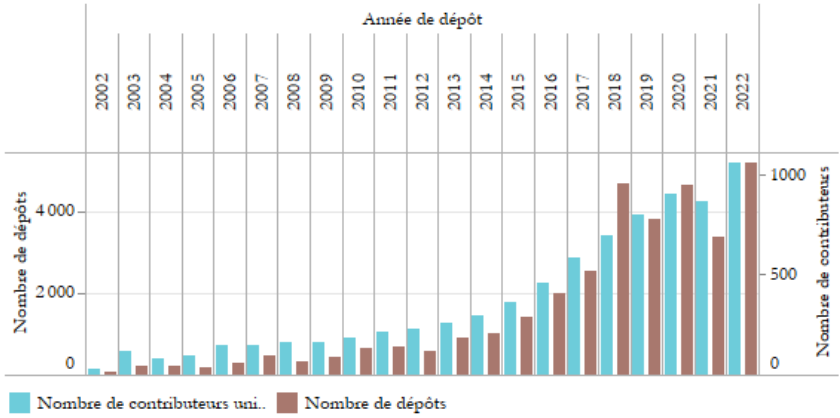

Figure 10. *Évolution du nombre de dépôts et du nombre de contributeurs relevant des SIC*

Historiquement majoritaires et parmi les premiers en SHS (plus de 50% des contributeurs pour 30% des dépôts jusqu'à 2004) à déposer en archive (fig. 11), l'évolution du nombre de dépôts et du nombre de contributeurs en SIC (fig. 10) n'est pas aussi constante qu'en SHS. Le nombre de dépôts annuels croît jusqu'en 2018 puis fluctue du fait de la crise sanitaire. Le nombre de contributeurs annuels, lui, est en constante augmentation à l'exception de l'année 2021. Bien que l'on ne distingue pas aussi nettement les périodes délimitées précédemment, l'accélération de la croissance dans ces deux paramètres commence dès 2014 et l'impact de la crise sanitaire apparaît plus marqué par un phénomène de nombre (restreints à la seule discipline SIC il y a de fait moins de contributeurs et sur le graphique les courbes de production sont moins lissées). Nous notons toutefois que l'accélération précoce, dès 2002, trouverait son explication par la présence de spécialistes de la documentation dans la discipline. La figure 11 détaille la proportion des contributeurs en SIC relativement aux SHS. Les deux premières années se distinguent : plus de 60% de tous les contributeurs en SHS les appartiennent aux SIC et déposent à eux seuls plus de 30% du total des

dépôts SHS. Ils seront encore plus de 40% la troisième année. Dès que le phénomène du dépôt en archive s'est répandu (2005) la répartition montre un important rééquilibre. La figure 12 entérine ces ratios, et montre qu'en 2020 les SIC comptent 900 contributeurs pour 10 000 dans toutes les autres disciplines des sciences humaines et sociales confondues.

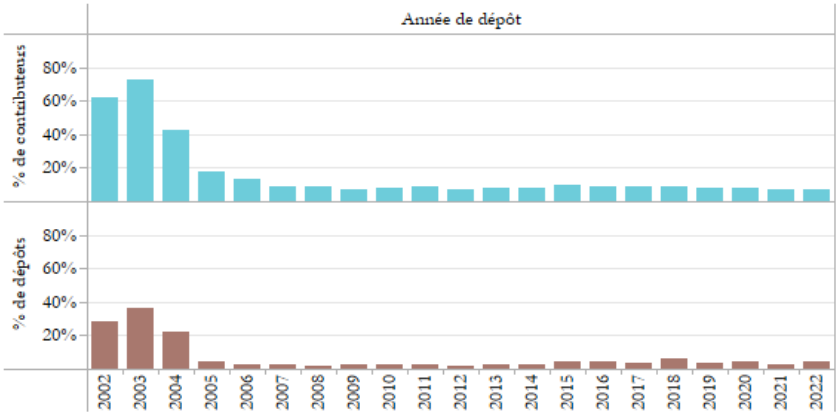

Figure 11. *Évolution de la proportion des contributeurs (haut) et des dépôts (bas) relevant des SIC par rapport aux SHS*

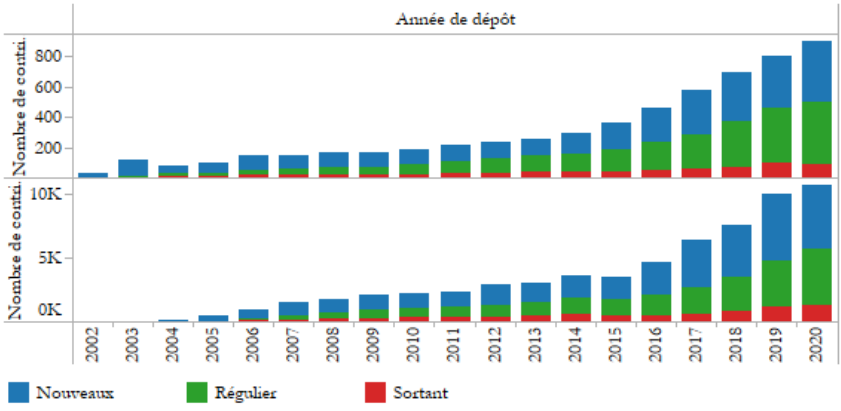

Figure 12. *Évolution des flux de contributeurs appartenant aux SHS (haut) et appartenant aux SHS (SIC exclues, bas)*

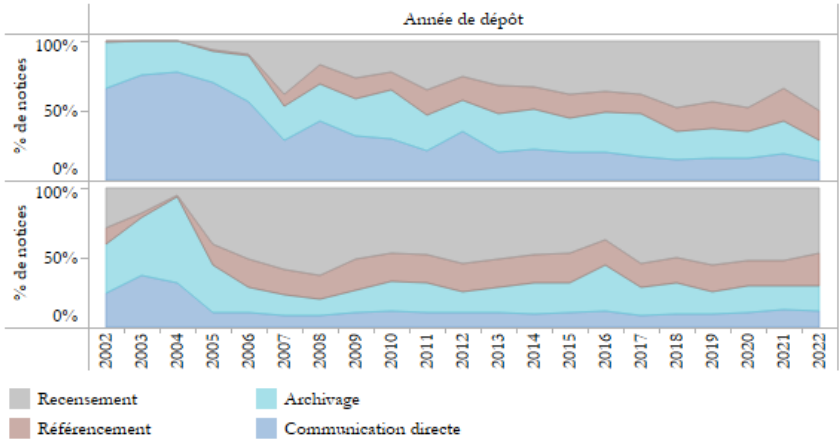

Figure 13. *Évolution des logiques de dépôt pour les notices relevant des SIC (haut) et relevant des SHS (SIC exclues, bas)*

L'évolution des logiques de dépôt (fig. 13) montre une singularité majeure des SIC relativement aux sciences humaines et sociales. Ainsi, les parts des dépôts effectués selon une logique de communication scientifique directe ou d'archivage, qui représentaient la majorité des dépôts jusqu'en 2014, n'ont cessé de décroître pour atteindre, respectivement, environ, 14% et 15% en 2022.

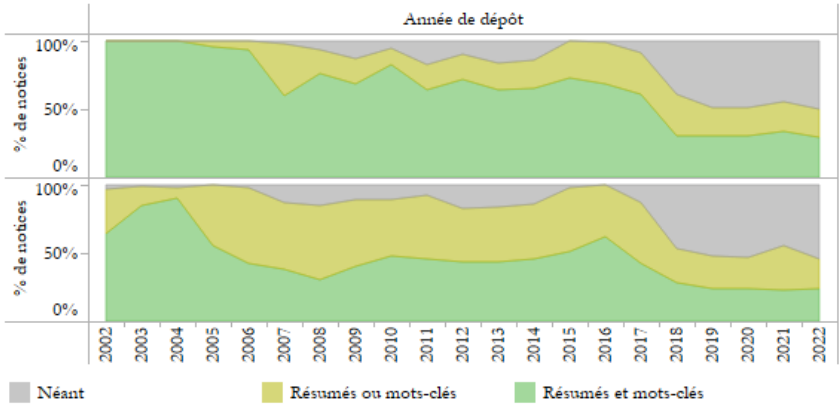

Figure 14. *Évolution du niveau de description pour les notices relevant des SIC en haut et celles relevant des SHS (SIC exclues) en bas*

Nous observons également une nette baisse dans la qualité de description des notices déposées (fig. 14) ; alors que, jusqu'en 2017 et à contrario des autres disciplines des SHS, les notices étaient plus décrites. Ainsi chaque année, au moins 60% des notices déposées étaient décrites par des résumés et des mots-clés (contre 30% seulement pour les SHS). La régression dans la qualité de description se produit plus tardivement, en 2018. L'évolution se stabilise par la suite avec, environ, 50% des

notices présentant un résumé ou des mots-clés en 2022, au même niveau que le reste des SHS.

Nous constatons que la pratique de l'archivage au sein des SIC se différencie de son groupement parent sur plusieurs points. Jusqu'en 2016, on observe un nombre de contributeurs plus importants que pour les autres disciplines des SHS. Ces contributeurs déposent plus souvent en joignant un texte intégral à la notice et renseignent mieux les métadonnées de ces dernières. Le lancement d'@rchiveSIC (Gallezot, 2003) en 2002 semble avoir impulsé une dynamique positive au sein de la discipline portée par les spécialistes de la documentation mais dont les effets s'effacent progressivement avec l'arrivée de nouveaux contributeurs

5.3 Impact des pratiques d'archivage sur le processus de diffusion

Nos résultats montrent qu'en SHS (tableau 2) ainsi qu'en SIC (tableau 3), mieux une notice est décrite, selon une dimension de complétude, plus elle est consultée. En SHS, mieux une notice est décrite, plus le texte intégral est téléchargé depuis l'archive HAL. Les corrélations sont toutefois moins fortes en SIC qu'en SHS.

	Résumé	Mots-clés	Fichier
Vues	$T \approx 32{,}74$, $p \approx 0$	$T \approx 17{,}95$, $p \approx 0$	$T \approx 29{,}10$, $p \approx 0$
Téléchargements	$T \approx 9{,}96$, $p \approx 0$	$T \approx 3{,}87$, $p \approx 0$	

Tableau 2. *Résultats des tests statistiques pour les échantillons de notices en SHS*

	Résumé	Mots-clés	Fichier
Vues	$T \approx 14{,}86$, $p \approx 0$	$T \approx 8{,}45$, $p \approx 0$	$T \approx 14{,}53$, $p \approx 0$
Téléchargements	**$T \approx 2.58$, $p \approx 0{,}01$**	**$T \approx 1{,}82$, $p \approx 0{,}7$**	

Tableau 3. *Résultats des tests statistiques pour les échantillons de notices en SIC*

Nos résultats ne nous permettent pas d'affirmer que le niveau de description d'une notice est corrélé avec le nombre de citations. Ce résultat peut toutefois s'expliquer par la démocratisation des sites pirates, comme, par exemple SciHub (Maddi & Sapinho, 2022) et l'incomplétude des bases de citation dans nos disciplines.

6 Discussion

Nos résultats montrent une adoption croissante de l'archivage de la production scientifique sur l'archive institutionnelle par la communauté scientifique française. Dans notre périmètre de recherche, nous constatons que le processus de diffusion du savoir est impacté par l'acte de dépôt d'une ressource scientifique sur l'archive électronique HAL. Les résultats présentés ici confirment également que le processus de dépôt tend, de plus en plus, à être négligé, questionnant alors les fonctions élémentaires de l'archive : au-delà du dépôt du texte intégral, les métadonnées descriptives les plus élémentaires font souvent défaut. Ce processus est d'autant plus négligé par les nouveaux contributeurs que par les contributeurs réguliers. Ces derniers sont plus habitués à l'utilisation du dispositif et, aussi, probablement plus sensibles aux enjeux de la science ouverte. Ceci montre que l'oubli de l'apposition de métadonnées riches à la description de sa production scientifique est probablement dû à un manque de connaissance sur les processus d'indexation et de recherche documentaire.

Nous avons constaté qu'il existe de grandes disparités disciplinaires, résultantes de politiques locales ou dynamisées par des initiatives intra-disciplinaires. Nous décrivons ici le cas des SIC où la dynamique historique impulsée par le lancement

d'ArchiveSIC a influé de manière positive sur la pratique d'archivage pendant plusieurs années suivant la création tend aujourd'hui à disparaître.

L'accompagnement des chercheurs vers une meilleure compréhension des enjeux de la science ouverte et de l'indexation tout en tenant compte des dissemblances entre sciences, apparaît comme incontournable pour augmenter la qualité des dépôts par le degré de remplissage des métadonnées : complétude et normalisation via des vocabulaires contrôlés dans l'idéal. Ceci permettrait, en conséquence, une meilleure diffusion des savoirs, développant ainsi de nouveaux usages tout en améliorant ceux existants. Par la suite, l'étude de l'évolution récente de l'interface du processus de dépôt sur l'archive HAL, qui peut constituer un premier élément de réponse à la problématique de l'appauvrissement des données, à travers les méthodes décrites ici, nous semble intéressante.

7 Références

Alhuay-Quispe, J., Quispe-Riveros, D., Bautista-Ynofuente, L., & Pacheco-Mendoza, J. (2017). Metadata quality and academic visibility associated with document type coverage in institutional repositories of Peruvian universities. Journal of Web Librarianship, 11(3-4), 241-254.

Article 30—LOI n° 2016-1321 du 7 octobre 2016 pour une République numérique (1)—Légifrance. (2016). https://www.legifrance.gouv.fr/jorf/article_jo/JORFARTI000033202841

Beall, J. (2005). Metadata and data quality problems in the digital library. Journal of Digital Information, 6(3).

Berthaud, C., Charnay, D., & Fargier, N. (2021). Diffuser et pérenniser le savoir scientifique : 20 ans d'histoire de HAL. Histoire de la recherche contemporaine. La revue du Comité pour l'histoire du CNRS, Tome X-n°2, Article Tome X-n°2. https://doi.org/10.4000/hrc.6330

Boullier, D. (2017). Pour des sciences sociales de troisième génération (SS3G). In P.-M. Menger & S. Paye (Éds.), Big data et traçabilité numérique (p. 163-184). Collège de France. https://doi.org/10.4000/books.cdf.5011

Broudoux, E. (2015). Contours du document numérique connecté. Documents et dispositifs à l'ère post-numérique., 7-15. https://archivesic.ccsd.cnrs.fr/sic_01327851

Dates clés de la science ouverte. (2023, mars 1). CNRS | Science Ouverte. https://www.science-ouverte.cnrs.fr/dates-cles-science-ouverte/

Gallezot, G. (2003). ArchiveSIC, Archive Ouverte en Sciences de l'Information et de la Communication : Rôle, fonctionnement et usage. Archimag.com - Guide pratique, XX. https://archivesic.ccsd.cnrs.fr/sic_00000600

Jaffe, R. (2020). Rethinking Metadata's Value and How It Is Evaluated. Technical Services Quarterly, 37(4), 432-443. https://doi.org/10.1080/07317131.2020.1810443

Larrieu, M., & Schöpfel, J. (2022, juin). Différences disciplinaires en contexte de Science ouverte. Étude avec les publications de l'archive ouverte HAL. 8ème conférence document numérique et société : Communication scientifique et science ouverte : opportunités, tensions et paradoxes. https://hal.science/hal-03760316

Maddi, A., & Sapinho, D. (2022). Does Open Access Really Increase Impact? A Large-Scale Randomized Analysis (arXiv:2206.06874). arXiv. https://doi.org/10.48550/arXiv.2206.06874

Mahé, A., & Prime-Claverie, C. (2017a). Qui dépose quoi sur Hal-SHS ? Pratiques de dépôts en libre accès en sciences humaines et sociales. Revue française des sciences de l'information et de la communication, 11, Article 11. https://doi.org/10.4000/rfsic.3315

Mahé, A., & Prime-Claverie, C. (2017b). Science ouverte et présence numérique des chercheurs en sciences humaines et sociales. Une étude exploratoire à partir de deux plateformes en ligne : HAL-SHS et Hypotheses.org. Document numérique, 20(2-3), 79-96. https://doi.org/10.3166/dn.2017.00010

MESRI (Éd.). (2019). Le Plan national pour la science ouverte : Les résultats de la recherche scientifique ouverts à tous, sans entrave, sans délai, sans paiement. 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-larecherche-scientifique-ouverts-a-tous-sans-entrave-sans-delai-sans-paiement.html

MESRI (Éd.). (2021). Le Plan national pour la science ouverte 2021-2024 : Vers une généralisation de la science ouverte en France. Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation, Paris. https://www.enseignementsup-recherche.gouv.fr/fr/le-plan-national-pour-la-science-ouverte-2021-2024-vers-une-generalisation-de-la-science-ouverte-en-48525

Morel-Pair, C. (2005). Panorama : Des métadonnées pour les ressources électroniques. Ateliers des réseaux de la documentation scientifique, Arcachon.

Park, J.-R. (2009). Metadata quality in digital repositories : A survey of the current state of the art. Cataloging & classification quarterly, 47(3-4), 213-228.

Remy, P. (2021). Name Dataset. In GitHub repository. GitHub. https://github.com/philipperemy/name-dataset

Reymond, D. (2022, février 24). SoViSu : Visibilité et lisibilité en SO. Séminaire IMSIC AXE 1 Brevets, captas et Science Ouverte.

Schöpfel, J., Chaudiron, S., Jacquemin, B., Kergosien, E., Prost, H., & Thiault, F. (2023). The Transformation of the Green Road to Open Access (No 2023020268). Preprints. https://doi.org/10.20944/preprints202302.0268.v1

Swan, A. (2010, février). The Open Access citation advantage : Studies and results to date [Monograph]. s.n. https://eprints.soton.ac.uk/268516/

Tabariès, A. (2022). Vers une métrique pour évaluer les métadonnées de documents scientifiques. Revue française des sciences de l'information et de la communication, 24. https://doi.org/10.4000/rfsic.12258

Visibilité des dépôts HAL : Moissonnage, signalement. (2023, mars 1). HAL Documentation. https://doc.archives-ouvertes.fr/guide_utilisateurs/visibilite-des-depots-hal-moissonnage-signalement/