

Traçage des données dans la recherche : Agrégation et utilisation ou vente des données d'usage par les éditeurs scientifiques

Jean-François Nominé, Porquet Thomas

▶ **To cite this version:**

Jean-François Nominé, Porquet Thomas. Traçage des données dans la recherche : Agrégation et utilisation ou vente des données d'usage par les éditeurs scientifiques : Une synthèse du Comité des services de bibliothèques scientifiques et des systèmes d'information de la Deutsche Forschungsgemeinschaft (DFG, Fondation allemande pour la recherche) 28 Octobre. 2022, 10.13143/dnzb-ym48 . hal-03642035

HAL Id: hal-03642035
https://cnrs.hal.science/hal-03642035

Submitted on 14 Apr 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Traçage des données dans la recherche : Agrégation et utilisation ou vente des données d'usage par les éditeurs scientifiques

Nominé Jean-François, Porquet Thomas

▶ **To cite this version:**

Nominé Jean-François, Porquet Thomas. Traçage des données dans la recherche : Agrégation et utilisation ou vente des données d'usage par les éditeurs scientifiques : Une synthèse du Comité des services de bibliothèques scientifiques et des systèmes d'information de la Deutsche Forschungsgemeinschaft (DFG, Fondation allemande pour la recherche). 2022, 10.13143/dnzb-ym48 . halshs-03634551v2

HAL Id: halshs-03634551
https://halshs.archives-ouvertes.fr/halshs-03634551v2

Submitted on 12 Apr 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

TRAÇAGE DES DONNÉES DANS LA RECHERCHE



07/04/2022

Agrégation et utilisation ou vente des données d'usage par les éditeurs scientifiques

Une synthèse du Comité des services de bibliothèques scientifiques et des systèmes d'information de la Deutsche Forschungsgemeinschaft (DFG, Fondation allemande pour la recherche)

Avertissement sur la traduction

Le présent document est la version française de la note d'information suivante :

> *DFG-Committee on Scientific Library Services and Information Systems. (2021). Data tracking in research: aggregation and use or sale of usage data by academic publishers. A briefing paper of the Committee on Scientific Library Services and Information Systems of the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation).*

publiée le 28 octobre 2021 et disponible sur Zenodo. https://doi.org/10.5281/zenodo.5937995

La traduction a été assurée par Jean-François Nominé, Inist-CNRS en association avec Thomas Porquet du consortium Couperin.org pour la révision. Cette traduction a été réalisée à la demande du consortium Couperin.org.

Comme certaines notes l'indiquent, la traduction de l'anglais a pu s'inspirer de l'original allemand là où plus de précisions ont été nécessaires. Le lecteur trouvera l'original sur le site de la DFG :
https://www.dfg.de/download/pdf/foerderung/programme/lis/datentracking_papier_de.pdf

DOI de la présente version: 10.13143/dnzb-ym48

Conformément à son original anglais, la présente traduction est placée sous licence Creative Commons Attribution 4.0 International

Le cliché en page de couverture de la présente version provient de la Deutsche Fotothek :

http://www.deutschefotothek.de/documents/obj/71820834/adf_rh_0000120

Conformément aux prescriptions de ce site, celui-ci est utilisable sous licence Creative Commons CC-BY-SA 3.0.

couperin.org
Consortium Unifié des Etablissements Universitaires et de Recherche pour l'Accès aux Publications Numériques

cnrs

TABLE DES MATIÈRES

1. DESCRIPTION DE LA SITUATION ACTUELLE 2

2. LA TRANSFORMATION DES GRANDS ÉDITEURS ET LEUR RELATION AVEC LA COMMUNAUTÉ DE RECHERCHE 5

2.1 Conséquences de la transformation des éditeurs en entreprises d'analyse de données 7

3. TYPES D'EXPLORATION DE DONNÉES 9

3.1 Données provenant de tiers grâce au microciblage 10

3.2 Données *bidstream* et analyse des ports 11

3.3 Les "logiciels espions" 12

4. CONCLUSION 13

La présente synthèse préparée par le Comité des services des bibliothèques scientifiques et des systèmes d'information (AWBI) de la Deutsche Forschungsgemeinschaft (DFG, Fondation allemande pour la recherche) porte sur le traçage des données dans les ressources électroniques de recherche et décrit les possibilités de pistage numérique des activités de recherche. Elle décrit comment les éditeurs scientifiques deviennent des spécialistes de l'analyse de données, en montre les conséquences pour la recherche et ses institutions, et identifie les types d'exploitation des données pratiqués. En tant que telle, elle a avant tout pour but de présenter les pratiques en cours en vue de favoriser la discussion dans la perspective que des positions puissent être adoptées quant aux conséquences de ces pratiques pour la communauté académique. Elle s'adresse à tous les acteurs du paysage de la recherche.

1. DESCRIPTION DE LA SITUATION ACTUELLE

Ces dernières années, différents types de marchés de données numériques sont apparus, que l'on peut classer en trois domaines : public, académique et commercial.[1] Dans le champ académique, outre les évolutions très positives telles qu'une amélioration de leur prise en charge, des règles juridiques concernant les données de recherche ainsi que l'amélioration de leur utilisation, d'autres évolutions doivent être examinées en détail et, si nécessaire, soumises à un contrôle réglementaire. Ce sont ces évolutions qui sont présentées dans ce qui suit. Un effet potentiellement négatif pour la communauté de la recherche résulte du mélange d'intérêts académiques et commerciaux ainsi que des lacunes réglementaires et des différences de législations sur le plan international.

Depuis quelque temps, les grands éditeurs scientifiques ont modifié en profondeur leur modèle commercial, avec des conséquences importantes pour la recherche : l'agrégation et la réutilisation ou la revente des traces d'usage par des utilisateurs sont devenues des aspects significatifs de leur activité.[2] Certains éditeurs se considèrent désormais explicitement comme des entreprises spécialisées dans l'analyse de l'information.[3] Leur modèle d'affaires bascule de la fourniture de contenu

[1] Putnings, M., "Datenmarkt", in : *Praxishandbuch Forschungsdatenmanagement*, 2021, p. 143, Praxishandbuch Forschungsdatenmanagement (degruyter.com)
[2] Aspesi, C., Allen, N. S., Crow, R., Daugherty, S., Joseph, H., McArthur, J. T., & Shockey, N., SPARC Landscape Analysis, 2019, 29 mars, https://doi.org/10.31229/osf.io/58yhb
[3] Voir, par exemple, la présentation que la société Elsevier dresse d'elle-même sur son propre site : "Elsevier est une entreprise mondiale d'analyse de l'information qui aide les institutions et les professionnels à améliorer les performances dans les domaines de la santé et de la science pour le bien de l'humanité" https://www.elsevier.com/de-de/about

à l'analyse de données. Cela induit le traçage - c'est-à-dire l'enregistrement et le stockage - des données d'utilisation que génèrent les chercheurs (c'est-à-dire les profils personnels, les données d'accès et d'utilisation, le temps passé sur une ressource, etc.) lorsqu'ils utilisent des systèmes d'information, comme c'est le cas lors de recherches documentaires, par exemple. Ce pistage scientifique est effectué avec un ensemble d'outils qui va du suivi des visites de sites via les dispositifs d'authentification jusqu'à des données détaillées en temps réel sur le comportement informationnel des personnes et des institutions. L'enregistrement d'informations telles que les pages visitées, les accès, les téléchargements - y compris le fait de recueillir des profils très précis sur le comportement au plan scientifique – s'opère parfois sans que les utilisateurs en soient suffisamment informés. Les données provenant de différentes sources peuvent être agrégées et combinées avec des informations personnelles supplémentaires, y compris des détails tirés de de la sphère non-académique.

Les éditeurs collectent ces données pour deux raisons : premièrement, dans l'objectif d'exploiter un nouveau secteur d'activité qui permet de tirer avantage sur le plan commercial des données sur le savoir, les évolutions de la recherche et les acteurs concernés. Deuxièmement, l'objectif est d'élargir la gamme des services offerts par les grands éditeurs. Les données peuvent être utilisées pour améliorer les services existants. Par exemple, les chercheurs peuvent recevoir automatiquement des suggestions de lecture ciblées et des références pointant vers des résultats de recherche dans leur domaine en fonction de leur profil personnel. Il est également possible de développer des services nouveaux. Outre la fourniture et la gestion des résultats de recherche sous la forme de littérature scientifique, des prestations sont également proposées qui touchent de plus en plus souvent à la gestion des données et aux logiciels de recherche.

Ces différents services pourraient être reliés les uns aux autres, ce qui les rendrait faciles d'utilisation pour les chercheurs. Pour de nombreuses activités du cycle de recherche, voire leur totalité, des chercheurs peuvent utiliser les services d'un unique fournisseur, qui peut en outre offrir aux institutions des services particuliers (par exemple, des systèmes d'information sur la recherche). Par exemple, RELX - la société mère à laquelle appartient Elsevier – distribue et installe le logiciel de systèmes d'information sur la recherche PURE dans des universités du monde entier,

et souligne explicitement qu'il est capable d'offrir une visibilité sur l'ensemble du cycle de recherche[4,5].

Cette évolution présente le risque d'empiéter significativement sur l'anonymat des chercheurs que la législation sur la protection des données garantit fondamentalement, rendant ainsi les institutions de recherche conjointement responsables de la violation du droit à l'autodétermination informationnelle. Le traçage numérique peut aussi encourager l'utilisation abusive des données et l'espionnage académique et entraîner une discrimination personnelle à l'encontre des chercheurs. Au vu de la jurisprudence[6] de la Cour suprême fédérale allemande, de l'arrêt Schrems II et du projet de l'Europe de statuer sur les plates-formes (législation sur les marchés numériques)[7], les organismes de recherche devraient prendre position sur ces pratiques.

La stratégie en matière de données récemment présentée par le gouvernement fédéral allemand n'aborde pas spécifiquement cette situation, mais elle mentionne le problème dans son principe - à savoir la monopolisation[8] croissante, l'abus de position dominante et l'utilisation illicite de données : "Dans l'utilisation des données, tout ce qui est techniquement possible n'est pas éthiquement justifiable et politiquement souhaitable."[9]

Dans l'ensemble, les chercheurs sont confrontés à la difficulté de trouver un équilibre entre la possibilité de profiter de la facilité que représentent des offres de services

[4] Voir la description du service à l'adresse suivante : www.elsevier.com/solutions/pure

[5] Elsevier a déclaré le 22 septembre 2021 : "PURE est un outil que les clients institutionnels utilisent pour traiter leurs données. Ces données restent la propriété du client, et lorsque le contrat prend fin, les clients récupèrent leurs données (avec l'hébergement dans le cloud) ou elles restent chez le client (sur site). Elsevier n'acquiert aucun droit sur ces données et ne les utilise à aucune fin. Si le client le souhaite, Elsevier n'aura pas accès à une installation PURE en particulier, même si elle est hébergée dans le *cloud*. PURE et, en fait, tous les produits et services d'Elsevier sont conformes au RGPD. Le logiciel offre toutes les possibilités de traiter des données personnelles dans le respect du RGPD s'il est configuré correctement. À compter de l'été 2021, aussi bien le partenaire d'hébergement d'Elsevier, Amazon, que PURE sont certifiés conformément à la norme ISO 27001. En outre, Elsevier conclut avec ses clients des accords de traitement des données qui tiennent compte des prescriptions en local. Un solide accord de protection des données est en place avec chaque université lorsqu'Elsevier sert ses clients de cette manière. Les données que les clients placent dans PURE sont isolées de celles des autres clients de PURE et de l'utilisation d'Elsevier en général, sauf si l'université choisit d'activer les fonctions de partage des données qu'elle contrôle. Comme pour tous les produits logiciels, les données instrumentales indiquent à Elsevier si notre logiciel fonctionne correctement."”"

[6] Note de traduction : la version allemande originale a permis d'éclairer le sens du terme *jurisdiction* employé dans la version anglaise traduite ici. En effet, le terme « Rechtschreibung » est moins ambigu.

[7] Par exemple, la législation sur les marchés numériques énonce explicitement l'objectif que les données collectées doivent servir non seulement aux intermédiaires mais aussi à promouvoir la concurrence et l'intérêt public.

[8] Datenstrategie der Bundesregierung, Kabinettsfassung en date du 27 janvier 2021, p. 21, Datenstrategie der Bundesregierung und die Ausschreibung des Bundesministeriums für Bildung und Forschung für Datentreuhandmodelle in den Bereichen Forschung und Wirtschaft du 08.01.2021, Bekanntmachung - BMBF.

[9] Datenstrategie der Bundesregierung, Kabinettsfassung du 27 janvier 2021, p. 7, Datenstrategie der Bun- desregierung.

groupées et la capacité de conserver la maîtrise de leurs données. Dans de nombreux cas, les chercheurs ne sont pas conscients de l'importance que représentent les données tirées de leurs activités et de la manière dont elles sont utilisées à des fins commerciales. La relation entre le monde de la recherche et les éditeurs doit être analysée avec cet objectif de parvenir à un équilibre entre ces deux pôles de la commodité et de la maîtrise. Pareille entreprise ne peut toutefois réussir que si elle s'appuie sur des règles juridiques claires qui garantissent un niveau élevé de transparence et la participation[10] de la communauté académique.

2. LA TRANSFORMATION DES GRANDS ÉDITEURS ET LEUR RELATION AVEC LA COMMUNAUTÉ DE RECHERCHE

Les éditeurs ont commencé à intégrer des solutions d'authentification à l'échelle individuelle et de suivi des utilisateurs dans leurs services il y a quelque temps déjà. Ce faisant, ils sont en mesure d'offrir des services techniques exclusifs pour l'ensemble du processus de recherche et l'analyse de données liées à la recherche. On peut citer comme exemple le contrat passé en 2020 avec Elsevier aux Pays-Bas, qui couvre des prestations décrites comme des services professionnels ainsi que la collecte de données personnelles.[11] Certains éditeurs intègrent également dans leur offre les stratégies[12] SeamlessAccess ou GetFTR, qui visent à permettre aux principaux fournisseurs de la recherche de proposer des informations d'une manière aussi close sur elle-même que possible, en s'appuyant sur une authentification simple et unique (Single Sign-On).[13] GetFTR et SeamlessAccess proposent des informations sur la nature des données qu'ils collectent et sur la manière dont ils prennent en compte la vie privée des utilisateurs.[14,15] Après un premier retour critique de la part des bibliothécaires, des adaptations ont été effectuées.[16]

[10] Note de traduction : le terme reflète l'emploi de « participation » dans le texte anglais qui, lui, traduit le terme allemand très spécifique de « Mitbestimmung », vocable qui est également employé dans le contexte de la concertation dans les entreprises, particulier en Allemagne, et qui est lui aussi restitué par le terme de « cogestion » dans un contexte de cet ordre. L'intention exprimée par le texte allemand semble plus forte sur ce point que sa restitution dans les autres langues en portant l'idée d'une association plus manifestement décisionnelle.

[11] Contrat signé UKB Elsevier SD 2020-2024.pdf (vsnu.nl). Nous nous référons en particulier à l'annexe 5, mais aussi à la section 7.6. du contrat.

[12] Voir https://www.getfulltextresearch.com et https://seamlessaccess.org

[13] Moore, S. A., "Individuation through infrastructure", in : Journal of Documentation 77(1) en date du 28 juillet 2020, https://doi.org/10.1108/JD-06-2020-0090

[14] GetFTR : GetFTR | Why GetFTR - GetFTR (getfulltextresearch.com) dans FAQ no.7 : https://www.getfulltextresearch.com/why-use-getftr/

[15] SeamlessAccess: Privacy and Trust - SeamlessAccess : https://seamlessaccess.org/about/trust/

[16] Hinchcliffe, L.J. : "Why are Librarians concerned about GetFTR?",, in : The Scholarly Kitchen,10 novembre 2019, https://scholarlykitchen.sspnet.org/2019/12/10/why-are-librarians-concerned-about-getftr ; Youngen, Ralph, Toler, Todd : "Lessons Learned : A Year with GetFTR", dans : The Scholarly Kitchen,16 février 2021, https://scholarlykitchen.sspnet.org/2021/02/16/guest-post-lessons-learned-a-year-with-getftr/

Les organismes de recherche allemands ont récemment signé des accords DEAL avec de grands éditeurs scientifiques (Springer Nature[17] et Wiley[18]) en vue d'assurer l'accès ouvert et des prix à l'avenant pour la fourniture et la publication des résultats de recherche.

Au moment de conclure des contrats avec des éditeurs, il est toujours important d'examiner de près les dispositions contractuelles touchant à la confidentialité des données et aux systèmes d'accès et d'authentification[19] qui seront utilisés. Par essence, l'accès le plus commode est celui qui ne nécessite aucune authentification, autrement dit un accès ouvert, même si, là encore, il est possible de suivre les usages via les plates-formes d'édition. Outre l'accès à la littérature, de nombreuses institutions sont également liées à un logiciel particulier, par exemple celui fourni par un fournisseur comme Elsevier.[20] Cette entreprise est également sous-traitante de la Commission européenne et effectue pour son compte la collecte de données sur la science ouverte (Open Science Monitor). [21]

Une si large gamme de services donne la possibilité de capter des informations sur le plus grand nombre possible de phases du processus de recherche et de les commercialiser auprès de tiers : en fin de compte, cela place les éditeurs ou les entreprises capables de fournir aux chercheurs, aux milieux politiques, aux universités et à la société dans son ensemble les informations les plus complètes, basées sur des données en relation avec les activités scientifiques. Cela signifie également que les éditeurs deviennent indispensables à la gouvernance des organismes et des universités. On parle déjà de l'émergence d'un "supercontinent"[22] dans la fourniture de données de recherche et de renseignements sur la recherche. Certaines données sur cette activité peuvent servir autant la recherche en elle-même que les processus de gouvernance complexes qu'implique la recherche moderne. On parle de bonnes pratiques lorsque, par exemple, les règles touchant à la collecte,

[17] Kieselbach, S., Projekt DEAL - Springer Nature Publish and Read Agreement. 2020, https://doi.org/10.17617/2.3174351
[18] Sander, F., Herrmann, G., Hippler, H., Meijer, G., & Schimmer, R., Projekt DEAL - John Wiley & Son Publish and Read Agreement, 2019, https://doi.org/10.17617/2.3027595
[19] Stellungnahme des Deutschen Bibliotheksverbands "Empfehlungen zu Methoden zur Kontrolle des Zugriffs auf wissenschaftliche Informationsressourcen", https://www.bibliotheksverband.de/fileadmin/user_upload/DBV/positionen/2019_11_26_Rundgespaech_RA21_-_Stellungnahme_Empfehlungen_final.pdf
[20] Cf. la liste des institutions qui utilisent le logiciel Pure, le système d'information sur la recherche d'Elsevier, https://www.elsevier.com/solutions/pure/clients
[21] Se reporter à : Microsoft Word - Open Science Monitor Methodological Note_April 2019.docx (europa.eu).
[22] Schonfeld, R.C. : " The Supercontinent of Scholarly Publishing? ", in : The Scholarly Kitchen, 3 mai 2018, https://scholarlykitchen.sspnet.org/2018/05/03/supercontinent-scholarly-publishing

l'utilisation et le partage des données sont transparentes et claires et que les données sont également disponibles à des fins non commerciales pour les acteurs au sein même des infrastructures de recherche (chez CrossRef par exemple).

Les conséquences de cette "organisation de la recherche pilotée par les données"[23], les conditions de sa réalisation et les structures qui la fournissent, la vendent et l'exploitent doivent en définitive être examinées et conçues par la recherche elle-même. Les organismes doivent également défendre la position que la collecte et l'utilisation des données – lorsque cela est nécessaire – doivent être non seulement légales, mais respecter aussi des valeurs éthiques telles que la transparence et la traçabilité, et s'appuyer sur le consentement en fournissant des informations complètes sur les conséquences, ainsi que sur d'autres aspects des bonnes pratiques en matière de données, en veillant à ce que ces pratiques constituent la base de tout accord passé avec les fournisseurs.

2.1 Conséquences de la transformation des éditeurs en entreprises d'analyse de données

Le risque existe que cette évolution du modèle commercial vers l'analyse des données entraîne une privatisation de la société de la connaissance et qu'en fin de compte, ce ne soit plus le secteur public, mais de plus en plus des entreprises privées qui jouissent d'un accès privilégié aux connaissances sur le contenu et les tendances de la recherche, ses institutions et ses acteurs. La recherche, en tant que bien public, se voit soumise à la logique de la privatisation des infrastructures et aux conséquences qui en découlent.[24] Ce modèle commercial concerne non seulement les grands éditeurs, mais aussi des fournisseurs de bases de données de recherche moins importants. Diverses études et initiatives - dont l'appel "The Cost of Knowledge" en 2012 ainsi que des organismes comme Science Europe[25] et des associations de bibliothèques - ont attiré à plusieurs reprises l'attention sur cette augmentation considérable du volume d'informations et de données détenues par des entreprises du secteur privé et sur le fait qu'une telle concentration de connaissances sur la

[23] Herb, U. : "Zwangsehen und Bastarde", in : Information. Wissenschaft & Praxis, 69 (2-3), 2018, p. 87.
[24] Barlösius, E., Infrastrukturen als soziale Ordnungsdienste. Ein Beitrag zur Gesellschaftsdiagnose. Francfort/M. 2019, chapitre 6.4 : " Infrastrukturierung der Forschung und infrastrukturierende Forschung ".
[25] "Science Europe plaide pour que les utilisateurs de données et l'utilisation à des fins de recherche soient clairement exclus du champ d'application de la législation sur les services numériques, afin d'éviter des effets involontaires sur les activités de recherche. Un acte législatif qui vise à lutter contre la vente de contenus illégaux sur de grandes plates-formes commerciales pourrait avoir des effets secondaires sur des secteurs d'intérêt public, sauf à introduire des exceptions appropriées." Science Europe, The Digital Services Act Should Not Have Unintended Effects on Research, 2020, https://www.scienceeurope.org/media/4s3bnhbr/20200908_se_response_dsa_consultation_final.pdf

recherche ne fait pas que profiter à l'innovation dans le seul domaine de la fourniture d'informations à la recherche.[26]

L'évolution esquissée ici d'une privatisation de l'industrie de la connaissance[27] s'oppose à la liberté de la recherche, au traitement des données personnelles prescrit par la loi et au droit de la concurrence. Plus précisément, le suivi des données non réglementé ou non détecté peut entraîner une violation de la liberté académique et de la liberté de recherche et d'enseignement ;

- constituer une violation du droit à la protection des données personnelles ;
- représenter une menace pour les chercheurs, car les données pourraient également devenir accessibles à des gouvernements étrangers et des régimes autoritaires ;
- attenter au droit de la concurrence, car les nouveaux participants n'ont guère de possibilités d'entrer sur le marché ;
- favoriser une dévalorisation des investissements publics dans la recherche, puisque les données sur l'activité de recherche peuvent être collectées par des concurrents commerciaux ou mises à leur disposition contre rémunération liée à l'espionnage industriel.

Les premiers cas de valorisation commerciale de données concernant les centres d'intérêt propres en recherche de scientifiques illustrent à quel point cette industrialisation de la connaissance au travers du pistage est déjà devenue une question critique.[28] LexisNexis, acteur international prestataire de solutions d'information et filiale du groupe RELX – dont Elsevier fait partie - a signé un accord pour livrer des données personnelles à l'ICE, l'agence américaine chargée de l'immigration et des douanes, pour un montant de 16,8 millions de dollars.[29] La situation se voit souvent aggravée par le fait que les établissements d'enseignement supérieur et les bibliothèques peuvent se rendre complices, à leur insu, de violations

[26] Par exemple Dobusch, L., "Kein Open-Access-Deal, dafür Spyware gegen Schattenbibliotheken", dans : netzpolitik.org, daté du 26 novembre 2020, https://netzpolitik.org/2020/neues-vom-grossverlag-elsevier-kein-open-access-deal-dafuer-mit-spyware-gegen-schattenbibliotheken/ ; die Stellungnahme des Deutschen Bibliotheksverbands "Empfehlungen zu Methoden zur Kontrolle des Zugriffs auf wissenschaftliche Informationsressourcen", www.bibliotheksverband.de/fileadmin/user_upload/DBV/positionen/2019_11_26_Rundgespaech_RA21_-_Stellung-nahme_Empfehlungen_final.pdf ; .

[27] Burgelman, J-C. : " Scholarly publishing needs regulation ", in : Research Professional News, 28 janvier 2021, https://www.researchprofessionalnews.com/rr-news-europe-views-of-europe-2021-1-scholarly-publishing-needs-regulation/

[28] Jung, J. : "UCLA School of Law Holds Contract with Companies Selling Personal Data to ICE", in : The Daily Bruin, 17 juillet 2020, https://dailybruin.com/2020/07/17/ucla-school-of-law-holds-contracts-with-companies-selling-personal-data-to-ice

[29] Biddle, S. : "LexisNexis to Provide Giant Database of Personal Data to ICE", in : The Intercept, 2 avril 2021, LexisNexis to Provide Giant Database of Personal Data to ICE (theintercept.com).

de la législation sur la protection des données, de la liberté académique et du droit de la concurrence. Les profils de données comportementales des personnels des universités allemandes peuvent être échangés et transférés de la même manière que celle qui a conduit à l'invalidation du Privacy Shield dans l'arrêt Schrems II, à savoir la transmission de données personnelles à un pays tiers hors de l'UE, puisque les mêmes parties prenantes sont impliquées.[30] En outre, des risques pourraient éventuellement survenir si les grands éditeurs présentaient un programme censuré sur le marché chinois. Le traçage pourrait conduire à la production de données personnalisées sur les personnes qui utilisent et recommandent les documents censurés, sans que les chercheurs concernés puissent déterminer qui a accès à ces données de pistage. Dans la perspective d'éventuelles modifications de la législation, Google a réagi récemment et annoncé un changement de sa politique de traçage ; ainsi, à l'avenir, il sera organisé de manière plus anonyme et sur la base d'une identification par "cohortes" plutôt que par utilisateurs individuels. [31]

3. TYPES D'EXPLORATION DE DONNÉES

Il existe trois types différents de fouille de données (*data mining*), c'est-à-dire des méthodes par lesquelles les données sont collectées et stockées par les éditeurs (données de tiers par le biais du microciblage, de données de flux d'enchères (*bidstream)*, et de l'analyse des ports (ou *port scanning)*, et des logiciels espions d'éditeurs), qui sont décrits *infra*. Il existe également différents outils : le portefeuille actuellement utilisé dans la recherche comprend des traqueurs de visites de pages, des outils de mesure d'audience où l'on peut agréger en profils différentes sources de données, le relevé d'empreintes numériques qui identifient même les utilisateurs qui veulent empêcher leur identification grâce aux paramètres du navigateur, et des outils de mise aux enchères en temps réel de données d'utilisateurs. Les outils de pistage sont pour la plupart produits par des prestataires tiers sous-traitants des grandes entreprises de l'internet, mais aussi par des entreprises spécialisées comme BlueKai, la plate-forme Big Data appartenant à Oracle, qui fait elle-même l'objet d'actions collectives en justice pour utilisation abusive de données personnalisées.[32] Comme

[30] Cf. Site du land de Basse-Saxe, Das SchremsII-Urteil des Europäischen Gerichtshofs und seine Bedeutung für Datentransfer in Drittländer, 2021, https://lfd.niedersachsen.de/startseite/themen/weitere_themen_von_a_z/internationaler_datenverkehr/das_schrems_ii_urteil_des_eugh_und_seine_bedeutung_fur_datentransfers_in_drittlander/das-schrems-ii-urteil-des-europaischen-gerichtshofs-und-seine-bedeutung-fur-datentransfers-in-drittlander-194085.html

[31] Neue Spielregeln: Warum Google Cookie-Tracking abschafft (netzpolitik.org)

[32] Lomas, N. : "Oracle and Salesforce Hit with GDPR Class Action Lawsuits Over Cookie Tracking Consent", in : TechCrunch, 14 août 2020, https://techcrunch.com/2020/08/14/oracle-and-salesforce-hit-with-gdpr-class-action-lawsuits-over-cookie-tracking-consent/

elles sont déjà associées institutionnellement à d'autres agrégateurs de données exploités par des services internet, ces données peuvent être compilées sous forme de profils pour être combinées avec d'autres encore et qui se rapportent à d'autres domaines de la vie des utilisateurs.[33] Les éditeurs ne divulguent pas jusqu'à quelle profondeur ils opèrent leurs traçages, si bien qu'à l'heure actuelle, nous ne pouvons que nous reporter à divers tests[34] montrant que toute personne qui accède à un article de la revue *Nature*, par exemple, est suivi par plus de 70 pisteurs.[35] Enfin, ajoutons que les outils utilisés peuvent être imparfaits, ce qui entraîne des conséquences encore plus dommageables pour les chercheurs sur le plan individuel.[36] Les trois principaux types de fouille de données mentionnés ici vont être brièvement décrits dans ce qui suit. On peut prévoir que les instruments de traçage numérique de la recherche seront continûment perfectionnés et leur application étendue, étant donné qu'ils procurent aux prestataires et aux entreprises des avantages concurrentiels considérables.

3.1 Données provenant de tiers grâce au microciblage

Le microciblage (*microtargeting*) consiste à viser des groupes cibles très spécifiques. Les éditeurs utilisent à la fois des données de première main et des données de seconde main. Les données de première main sont les traces directement laissées par des utilisateurs, tandis que les données de seconde main sont des données achetées, qui sont à leur tour synthétisées en profils de données précis par des tiers, principalement les grandes entreprises de l'internet. Les éditeurs ont mis en place une grande variété de ces *sources de données tierces* sur leurs plates-formes, qu'il s'agisse des *trackers* (pisteurs) très répandus utilisés par Google ou Facebook, de ceux utilisés par des fournisseurs tels que BlueKai et Krux Digital, d'outils d'empreintes numériques de navigateurs tels que Double Click ou d'outils d'agrégation de données d'audience d'Adobe, Neustar, Oracle, AddThis et autres. Le code JavaScript de ces prestataires tiers peut avoir accès au *Document Object Model* (DOM) du site web en question, de sorte qu'il est capable de lire et identifier le texte qui suscite l'attention de l'utilisateur, les documents vers lesquels il navigue

[33] Vogel, C. : "Kennen Sie Google CASA ?", in : Medinfo. Informationen aus Medizin, Bibliothek und Fachpresse, www.medinfo-agmb.de/archives/2020/07/08/6880

[34] "Digital Library Federation, Endangering Data. Interview with Sarah Lamdan", à voir sur www.diglib.org/endangering-data-interview-with-sarah-lamdan ou Lamdan, S. : "Social Media Privacy: A Rallying Cry to Librarians", in : The Library Quarterly 85 (3), 2015, p. 261-277 https://academicworks.cuny.edu/cl_pubs/52 ; les études de Wolfie Christl sur RELX et ThreatMetrix, https://twitter.com/wolfiechristl/status/1295655040741445632 et https://crackedlabs.org

[35] Brembs, B : https://twitter.com/brembs/status/1301897878387003398

[36] Voir par exemple Lamdan, S. : "Librarianship at the Crossroad of ICE Surveillance ", in : In the library with the lead pipe, 13 novembre 2019, et Swauger, S., https://twitter.com/SheaSwauger/status/1205587676172144641

ensuite et les mots-clés de recherche qu'il saisit sur la plate-forme. Étant donné que de nombreux fournisseurs font appel plus ou moins aux mêmes prestataires extérieurs ou qu'ils échangent parfois des données entre eux, le comportement informationnel des membres de l'enseignement supérieur peut être collecté sur plusieurs plates-formes et, dans le cas de Google, Facebook ou Twitter, il peut être relié à des renseignements déjà disponibles sur leurs autres activités sur les réseaux.[37] Dans le cas de fournisseurs tels que Acxiom/Liveramp, il leur est possible d'établir une synchronisation entre les activités en ligne et hors-ligne, puisque des données sont également disponibles sur les achats, les permis de conduire, la consommation télévisuelle, les listes électorales, les casiers judiciaires, etc. [38]

3.2 Données *bidstream* et analyse des ports

L'intégration de tierces parties est souvent critiqué et, dans certains cas, celle-ci n'est plus appuyée par les grandes entreprises et institutions de l'internet, de sorte que des alternatives comme la collecte de données *bidstream* (données d'enchères en temps réel) sont maintenant également employées, par un recueil en arrière-plan de données sur la localisation, les appareils et les données utilisées. Les données de l'utilisateur sont mises aux enchères en temps réel en même temps qu'une variété d'informations individuelles telles que la localisation, l'adresse IP, des renseignements sur son équipement et bien plus encore ; elles sont ensuite transmises et reliées à un identifiant qui caractérise les personnes en question de manière fiable, même sans cookie.[39] Le simple fait de rechercher des ports ouverts sur les ordinateurs et/ou les réseaux d'autrui afin d'y infiltrer des logiciels malveillants ou de surveillance, par exemple, frise l'illégalité au regard du droit allemand, car il peut être considéré comme une étape préliminaire à certaines infractions qu'il sanctionne (§§ 202c, 303b du Code pénal allemand - StGB). Cette méthode reste néanmoins largement utilisée, en partie à des fins de prévention de la fraude et en partie comme outil de traçage.

Un exemple qui a suscité l'intérêt du public est celui de ThreatMetrix, qui fait partie de LexisNexis Risk Solutions/RELX et prétend pouvoir identifier 4,5 milliards d'appareils. ThreatMetrix est installée sur ScienceDirect, la plate-forme par laquelle les chercheurs consultent le contenu des revues publiées par Elsevier. Les relations

[37] Hanson, C. : User Tracking on Academic Publisher Platforms, 2019, www.codyh.com/writing/tracking.html
[38] Cf. le diagramme dans Christl, W. : Corporate Surveillance in Everyday Life, p. 55, https://crackedlabs.org/dl/CrackedLabs_Christl_CorporateSurveillance.pdf
[39] Cf. Ryan, J. : Briefing on adtech, RTB, and the GDPR at dmexco Brave Event, Diapositive 45, www.slideshare.net/JohnnyRyan/briefing-on-adtech-rtb-and-the-gdpr-at-dmexco-brave-event.

du groupe RELX avec différentes administrations du gouvernement américain font déjà l'objet de pétitions publiques aux États-Unis.[40] Tant que les éditeurs ne divulguent pas leurs pratiques de *tracking*, on continue de supposer que les données recueillies à l'aide de ces logiciels de pistage peuvent être également utilisées dans d'autres produits de la branche Risk Solutions[41], par exemple dans le domaine des analyses proposées à des entreprises et des pouvoirs publics.[42]

3.3 Les "logiciels espions"

Proposés à des bibliothèques et assortis de rabais sur d'autres services, les "logiciels espions" sont employés dans le but de passer le pistage des chercheurs à l'échelle. Les "spywares" sont des logiciels additionnels destinés à être installés dans les bibliothèques pour collecter des données biométriques comme la vitesse de frappe ou le type de déplacements de la souris afin de caractériser les utilisateurs malgré le recours à des serveurs proxy et des tunnels VPN.[43] Les entreprises et les prestataires peuvent faire valoir que ces logiciels offrent la capacité d'attaquer judiciairement les utilisateurs de bibliothèques pirates.[44,45] Toutefois, ces logiciels espions compromettent la sécurité des réseaux des universités et peuvent exposer celles-ci à toutes sortes d'attaques. Par conséquent, on ne peut en recommander l'utilisation.[46]

---

[40] American Civil Liberties Union : ACLU Calls On Tech Companies to End Their Alliance with ICE and CBP, 2020, www.aclu.org/news/immigrants-rights/aclu-calls-on-tech-companies-to-end-their-alliance-with-ice-and-cbp.

[41] Risk & Business Analytics – RELX

[42] Cf. la documentation de Wolfie Christl à l'adresse suivant : https://twitter.com/wolfiechristl/status/1286341387718397952

[43] Cf. Mehta, G. : " Proposal to Install Spyware in Universities Libraries to Protect Copyrights Shocks Academics ", in : Coda, 13 novembre 2020, www.codastory.com/authoritarian-tech/spyware-in-libraries

[44] L'équivalent allemand - Schattenbibliotheken – calque de "shadow library" – est devenu un terme consacré et se trouve utilisé ici à titre d'exemple : Ball, R. : Wissenschaftskommunikation im Wandel, Springer, 2020, p. 127.

[45] Dans une version antérieure de ce document, l'ISP était citée. Cette citation a été supprimée dans la présente version en raison d'une déclaration de PSI du 9 juin 2021 indiquant ce qui suit:
"1. PSI ne travaille pas avec la SNSI à quelque titre que ce soit.
2. PSI ne trace pas les utilisateurs de bibliothèques clandestines.
3. PSI n'est en rien liée à, ou n'a connaissance de, quelque logiciel espion que ce soit.
4. PSI n'a pris part à aucune poursuite envers des "utilisateurs" de bibliothèques clandestines ni ne présume que pareils "utilisateurs" aient fait l'objet de poursuite de la part de qui que ce soit. "

[46] Dans une version antérieure de ce document, l'Initiative de sécurité des réseaux universitaires (SNSI) était mentionnée dans ce paragraphe. Cette mention a été supprimée suite à une déclaration de sa part en date du 8 septembre 2021 :
"La SNSI encourage vivement ses clients à appliquer des mesures de sécurité forte protégeant l'accès aux données des chercheurs et des étudiants et le contenu qui leur est fourni par ses membres, mais par ailleurs :

- elle n'emploie ni ne préconise d'utiliser des logiciels espions (par exemple, pour collecter des données biométriques telles que la vitesse de frappe ou le type de mouvement de la souris afin de pouvoir caractériser des utilisateurs malgré leur utilisation de serveurs proxy et de tunnels VPN) ;
- elle ne prend aucune disposition, ni ne préconise une quelconque incitation auprès des bibliothèques, favorable à l'installation et l'exploitation de logiciels espions en propre."

4. CONCLUSION

Pareils traçages de la recherche peuvent aller à l'encontre de la liberté académique et du droit de la personne à l'autodétermination informationnelle. Ils peuvent mettre en danger les chercheurs et entraver la liberté de concurrence dans le secteur de la fourniture d'information. C'est pourquoi les chercheurs et les institutions académiques doivent prendre conscience du problème et établir clairement les conditions juridiques, techniques et éthiques qui encadrent l'information qui leur est fournie - non seulement pour éviter de violer involontairement la législation en vigueur, mais aussi pour garantir que les personnels de la recherche soient correctement informés et protégés.

En publiant cette synthèse, le comité AWBI poursuit l'objectif de susciter un large débat au sein de la communauté académique - parmi les décideurs académiques, parmi les chercheurs comme au sein des infrastructures d'information au niveau institutionnel - afin de porter une réflexion sur les pratiques de traçage, leur légalité, les mesures nécessaires au respect de la protection des données, les conséquences qu'entraîne l'agrégation des données d'utilisation, et ainsi permettre l'adoption de pareilles mesures.

La collecte de données sur la recherche et l'activité de recherche peut avoir son utilité à condition de suivre des lignes de conduite claires et transparentes, de réduire les risques pour chaque chercheur à son niveau et de veiller à ce que les organisations de recherche disposent de la capacité d'utiliser ces données, si ce n'est d'en posséder la maîtrise.