Pistage et profilage des usagers par les grands éditeurs scientifiques : quels nouveaux enjeux pour les bibliothèques ?

Marie Nikichine, Porquet Thomas

couperin.org
Consortium unifié des établissements universitaires et de recherche pour l'accès aux publications numériques

DECRYPTAGE

Pistage et profilage des usagers par les grands éditeurs scientifiques : quels nouveaux enjeux pour les bibliothèques ?

Ce document propose une synthèse de travaux récents, en particulier les rapports de l'association SPARC et du Comité des services des bibliothèques scientifiques et des systèmes d'information (AWBI) de la Deutsche Forschungsgemeinschaft (DFG, Fondation allemande pour la recherche). Ces rapports documentent et nous alertent sur les graves menaces qui pèsent sur la protection de la vie privée des usagers de ressources documentaires, depuis qu'une part croissante d'éditeurs a cessé de se consacrer à la seule fourniture de documents (ouvrages, revues) pour s'orienter vers l'octroi de licences pour l'accès à des contenus en ligne qui restent hébergés sur leurs serveurs. Au cours de cette mutation, ces entreprises ont acquis un très grand contrôle dans la collecte, la conservation et l'analyse des données de lecture et de recherche des utilisateurs de la bibliothèque, y compris quand il s'agit de données à caractère personnel (ou DCP).

La transformation en cours : de l'édition à l'analyse de données

L'acquisition par les grands éditeurs des services et produits qui sous-tendent l'ensemble du cycle de vie de la recherche témoigne aujourd'hui d'orientations stratégiques nouvelles de la part de ces sociétés. Elles visent à intégrer l'ensemble des infrastructures scientifiques, et manifestent un intérêt croissant à collecter et monétiser les nombreux flux de données que leurs plateformes peuvent générer à chaque étape du parcours des utilisateurs : lecture, publication, recherche de financements, de poste, mise en relation avec des collègues, etc.

Le pistage des utilisateurs, qui serait impensable dans le cadre d'une bibliothèque physique, se fait désormais de façon routinière et intégrée dans des dispositifs techniques déployés massivement en ligne. Si des ouvrages et des revues en ligne sont encore vendus aux bibliothèques, la commercialisation des données utilisateurs et des analyses qui en sont faites constitue un secteur d'activité nouveau, radicalement différent de l'édition académique.

Cette dynamique est nommée « capitalisme de l'information », voire « capitalisme de la surveillance »[1], et ne concerne pas que les fournisseurs de bibliothèques mais un grand nombre d'entreprises, qui s'éloignent de la vente de biens physiques pour s'orienter vers la vente de données obtenues de plateformes numériques.

Dans ce paysage en mutation, Elsevier est un exemple typique : c'est l'un des plus grands fournisseurs de collections pour de nombreuses bibliothèques universitaires. Ses outils et services sont organisés pour couvrir la totalité du processus de production de connaissances. La société mère d'Elsevier, Reed Elsevier LexisNexis (RELX), se présente explicitement comme « fournisseur mondial d'outils d'analyse et de décision basés sur l'information pour les professionnels et les entreprises » à ses investisseurs et actionnaires[2]. L'activité "risque" de RELX, qui fournit des services aux entreprises, aux gouvernements et aux organismes chargés de l'application de la loi, s'appuie sur de vastes bases de DCP et dépasse sa division Elsevier en revenus et en rentabilité. RELX agit désormais comme courtier en données (data broker).

Des pratiques généralisées mais problématiques

Les pratiques en matière d'exploitation des données décrites ci-dessous sont adoptées par de nombreuses entreprises et organisations qui pistent et collectent les données des utilisateurs. Ces données alimentent des modèles commerciaux certes très rentables mais intrusifs sur le plan de la vie privée. Voici quelques exemples de techniques et pratiques actuellement en place :

- L'utilisation de marqueurs web, de cookies et d'autres méthodes invasives de surveillance du web, qui permettent de suivre le comportement de l'utilisateur au delà du seul site web de l'éditeur (traqueurs, relevés d'empreintes numériques, etc. 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 plateforme Big Data appartenant à Oracle)
- La collecte extensive d'un large éventail de DCP (par exemple, des données comportementales et de localisation) à partir du site de l'éditeur, combinée à des données recueillies auprès de sources extérieures
- La collecte de DCP par des tiers, y compris des moteurs de recherche, des plateformes de médias sociaux et d'autres agrégateurs de données personnelles et profileurs tels que Google, Adobe, Cloudflare et New Relic, par le biais d'une utilisation intensive des traqueurs tiers sur le site de l'éditeur.
- La fourniture de DCP à d'autres produits de l'éditeur ou à d'autres unités commerciales au sein d'un même groupe (y compris des produits et services « risque » vendus à des entreprises, des gouvernements et des organismes chargés de l'application de la loi), le traitement et la divulgation de DCP (et de DCP déduites d'autres DCP) à des fins de publicité et de marketing ciblés et personnalisés.

Les bibliothèques doivent donc être conscientes du fait que les grands acteurs de ce domaine pistent leurs usagers de la même manière que les plateformes de vente en ligne, les sites d'information et les réseaux sociaux.

Quelles garanties pour la protection de la vie privée des usagers ?

En plus du cadre légal déjà existant[3] et de son application concrète, la protection de la vie privée des utilisateurs dépend aussi de la qualité des protections garanties par les fournisseurs à leurs clients. Les modalités d'exploitation et d'usages des DCP (telles qu'elles sont par exemple négociées dans les contrats) figurent souvent dans les « politiques de confidentialité », mais sont parfois réparties dans différents documents distincts du contrat et nécessitent, pour s'en faire une idée précise, une expertise et un temps que beaucoup d'institutions n'ont pas. C'est l'utilisateur isolé qui se retrouve en première ligne, peu éclairé sur des clauses longues et complexes, auxquelles il va se soumettre, quand il n'est pas amené à y consentir malgré lui via des interfaces trompeuses (ou « dark patterns »).
Il faut désormais se préoccuper plus attentivement de la latitude accordée par ces clauses aux entreprises qui commercialisent ces produits. Les usagers et les institutions doivent se saisir sans attendre de ces enjeux, au vu de l'intégration croissante de ces pratiques de pistage à d'autres secteurs d'activité, y compris celui des outils d'analyse de la recherche et le courtage de données.

La concentration des acteurs du secteur de l'édition

Au cours de la dernière décennie, on a pu observer une tendance à la concentration du marché dans l'écosystème de la diffusion de l'information scientifique[4] : un oligopole de 6 groupes internationaux domine le marché mondial.
En 2021 et 2022, les fusions se sont accélérées entre de grands éditeurs et des entreprises associées :

- **Wiley** a acquis l'éditeur en accès ouvert **Hindawi**, la société de services d'édition **J&J Editorial**, celle de logiciels d'édition savante **eJournalPress** et la plateforme **Knowledge Unlatched**, modèle pour la publication de monographies en accès ouvert,
- **Springer Nature** a acquis **Atlantis Press**, un éditeur en accès ouvert,
- **Clarivate** a acquis **ProQuest**,
- **Copyright Clearance Center** a acquis **Ringgold**,

- **Elsevier** a acquis **Interfolio**, une entreprise de technologie universitaire dont le portefeuille comprend à la fois une plateforme de recherche d'emploi universitaire nommée **Dossier**, une plateforme d'analyse de recherche nommée **Researchfish** et la plateforme de données administratives **Faculty Information System (FIS).**

Ces regroupements montrent comment les grands éditeurs diversifient leurs revenus en vendant non seulement l'accès à des articles publiés, mais aussi en pistant les parcours et les interactions des usagers sur leurs plateformes, puis en vendant des analyses exploitant les données ainsi collectées. Les produits destinés à gérer les différents processus de pré-publication, de publication et d'évaluation de la recherche appartenaient auparavant à des entreprises indépendantes. Ils peuvent désormais être regroupés au sein d'entités uniques et intégrées. **Une part croissante du processus de recherche se trouve captée et monétisée par ces grands conglomérats.**
La tendance des grands éditeurs académiques à acquérir ou à fusionner avec des entreprises de données ne peut qu'être source de **préoccupations croissantes concernant le recueil et l'utilisation des données recueillies au moyen d'outils et de plateformes de recherche**. Les données collectées via ces plateformes propriétaires sont soumises aux politiques de confidentialité mises en place par l'entreprise, qui peuvent tout à fait prévoir leur revente à des tierces parties.
Au sein du groupe RELX, LexisNexis déclare vendre des produits basés sur des données personnelles à 70% des administrations locales et à près de 80% des agences fédérales aux États-Unis. Les revenus générés par cette activité ont atteint le même niveau que ceux de la branche éditoriale en 2022, et connaissent une progression plus forte[5]. Les données de LexisNexis sont aussi vendues à des sociétés de sélection de candidats à la location, à des systèmes de soins de santé, à des compagnies d'assurance et à des institutions bancaires. Elsevier fait partie du même groupe RELX, qui commercialise donc des DCP à toute une variété de clients, dont les bibliothèques ne sont plus qu'un type parmi d'autres.

Quelles stratégies pour les bibliothèques face à ces pratiques ?

La collecte et la divulgation de données sur l'identité d'une personne, sur l'endroit où elle se trouve et sur ce qu'elle recherche et lit, par la même entreprise qui fournit des produits sophistiqués de surveillance et de courtage de données aux entreprises et aux gouvernements, doivent nous alarmer. Utilisateurs, institutions, décideurs politiques et autorités de régulation doivent évaluer le danger de ces pratiques, qui interrogent sur les risques d'une propriété simultanée d'infrastructures universitaires clés et d'entreprises de surveillance et de courtage de données.
Au vu des évolutions du secteur de l'édition académique, les utilisateurs et les institutions doivent évaluer et traiter activement les risques potentiels pour la vie privée, pendant que cette transition est en cours, et avant qu'elle ne soit achevée.
L'enjeu des négociations que nous menons en tant que bibliothécaires dépasse la négociation de l'accès à des contenus. Il doit désormais être aussi indispensable :

- de négocier le contenu que les entreprises obtiennent dans le cadre de l'utilisation de leurs produits, et d'obtenir des conditions contractuelles plus strictes,
- de supprimer des contrats les clauses qui empêchent le partage des conditions d'exploitation des DCP et réduire ainsi les asymétries d'information avec les fournisseurs lors des négociations,
- de recalibrer les relations avec les fournisseurs dont les pratiques en matière de protection de la vie privée sont contraires aux attentes des bibliothèques et présentent des risques pour les utilisateurs.

C'est le chemin pris par nos collègues allemands dans leur dernier contrat DEAL-Elsevier, rendu public[6], qui comporte une section 7.6 sur la "Data Protection" dans laquelle il est question de :

- réduire au maximum la génération, le pistage, la collecte, le partage, le transfert et la revente de DCP sans le consentement éclairé des ayant-droits. Ces données à caractère personnel seront nécessairement traitées dans un pays de l'UE.

- faire mettre en place par Elsevier une plateforme de gestion des cookies sur leurs sites pour permettre aux ayant-droits de visualiser la nature et la durée des cookies, et de gérer leurs choix en ce qui concerne les cookies qui ne sont pas strictement nécessaires pour fournir, gérer, exploiter et sécuriser ces sites
- former un groupe de travail de 12 mois, pour discuter des mesures à prendre pour renforcer la protection des données et de la vie privée.

Conclusion

Les bibliothèques sont confrontées à l'évolution des plateformes d'accès à la documentation en des espaces clos et propriétaires dont les prix sont difficilement négociables et dans lesquels une large partie des publications ne sont ni accessibles ni exploitables.
En effet, dans un mouvement concomitant, les mêmes éditeurs qui vont tirer profit de l'usage d'outils d'analyse (et d'intelligence artificielle) sur les données qu'ils collectent, souhaitent faire inscrire dans les contrats des clauses qui restreignent, voire interdisent de procéder aux mêmes types de traitements par leurs abonnés.
Parce qu'elles ont affaire à des monopoles (ou des oligopoles dans le cas des revues académiques), les bibliothèques n'ont guère le choix des fournisseurs dont elles souhaitent utiliser les services. Pour autant, cela ne signifie pas qu'elles n'ont aucun moyen de faire face aux risques évoqués dans cette note : il faut qu'il soit à la portée de chaque professionnel de **prendre plus clairement conscience des risques liés aux pratiques en vigueur, d'intégrer cet aspect aux négociations documentaires, de partager les avancées obtenues** dans ce cadre, et d'alerter largement sur les enjeux.

Date : 26/01/2024

Rédaction :

- Marie Nikichine (Université Paul-Valéry Montpellier 3)
- Thomas Porquet (consortium Couperin)

Summary

This document summarizes recent work, in particular reports by the SPARC association and the Committee for Scientific Library Services and Information Systems (AWBI) of the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation). These reports document and alert us to the serious threats to the privacy of users of information resources, since a growing proportion of publishers have moved away from the sole supply of documents (books, journals) towards licensing access to online content that remains hosted on their servers. During this transition, these companies have acquired considerable control over the collection, storage and analysis of library users' reading and research data, including personal data.

Références

[1] Zuboff S., 2020, L'Âge du capitalisme de surveillance. Le combat pour un avenir humain face aux nouvelles frontières du pouvoir, trad. de l'anglais (États-Unis) par B. Fomentelli et A.-S. Homassel, Paris, Éd. Zulma

[2] "RELX is a global provider of information-based analytics and decision tools for professional and business customers" https://www.relx.com/our-business/our-business-overview

[3] Pour répondre aux évolutions du numérique, l'Union européenne s'est dotée d'un nouveau cadre juridique en matière de protection des données personnelles contenant en particulier le Réglement général sur la protection des données (RGPD) https://www.vie-publique.fr/dossier/19598-rgpd-protection-des-donnees-personnelles-ce-qui-change-en-mai-2018

[4] https://librarytechnology.org/mergers/

[5] Cf. le rapport annuel 2022 de RELX : https://www.relx.com/~/media/Files/R/RELX-Group/documents/reports/annual-reports/2022-ar-sections/relx-2022-overview.pdf

[6] Projekt DEAL – Elsevier Publish and Read Agreement https://pure.mpg.de/pubman/faces/ViewItemOverviewPage.jsp?itemId=item_3523659

Références complémentaires

- Breeding, M. (s. d.). History of Mergers and Acquisitions in the Library Technology Industry. Library Technology Guides. Consulté 20 décembre 2023, à l'adresse http://librarytechnology.org/mergers/
- 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). Zenodo. https://doi.org/10.5281/ZENODO.5937995
- Elsevier B.V., & MPDL Services gGmbH, Max Planck Society (2023). Projekt DEAL – Elsevier Publish and Read Agreement. https://doi.org/10.17617/2.3523659
- Lamdan, S., & Mann, S. (2023). What Does the Transition from Publishing to Data Analytics Mean for Libraries? NASIG Proceedings, 37(0). https://doi.org/10.3998/nasig.4026
- Nominé, J.-F., & Porquet, T. (2022). Traçage des données dans la recherche : Agrégation et utilisation ou vente des données d'usage par les éditeurs scientifiques. https://doi.org/10.13143/dnzb-ym48
- Posada, A., & Chen, G. (2018). Inequality in Knowledge Production: The Integration of Academic Infrastructure by Big Publishers. In L. Chan & P. Mounier (Éds.), ELPUB 2018: Vol. Connecting the Knowledge Commons: From Projects to Sustainable Infrastructure. Association Francophone d'Interaction Homme-Machine (AFIHM). https://doi.org/10.4000/proceedings.elpub.2018.30
- Yoose, B., & Shockey, N. (2023). Navigating Risk in Vendor Data Privacy Practices: An Analysis of Elsevier's ScienceDirect. SPARC. https://doi.org/10.5281/ZENODO.10078610