Ouverture des données et archives du web : enjeux, pratiques et limites

Dorothée Benhamou-Suesser, Fred Pailler, Valérie Schafer, Guillaume Garcia

Dans Terrains & travaux 2023/2 N° 43 , pages 281 à 294
Éditions ENS Paris-Saclay

ISSN 1627-9506

DOI 10.3917/tt.043.0281

Date de mise en ligne : 14/02/2024

Dorothée Benhamou-Suesser, Fred Pailler, Valérie Schafer

Ouverture des données et archives du web : enjeux, pratiques et limites

Entretien mené par Selma Bendjaballah et Guillaume Garcia

L'ENTRETIEN QUI SUIT offre un témoignage sur les enjeux, pratiques et difficultés suscités par l'ouverture et l'exploitation des données du web et des réseaux sociaux numériques. Si de plus en plus d'enquêtes mobilisent sites web, discussions sur des forums, échanges sur les réseaux sociaux, etc., de fait, une multitude de pratiques existent au sein des sciences humaines et sociales (SHS) : des chercheur·es font leur propre collecte, d'autres utilisent des collections déjà créées ; certain·es croisent des données issues de plusieurs plateformes, quand d'autres se concentrent sur un réseau social numérique ; certain·es exploitent les contenus textuels exclusivement, alors que d'autres abordent aussi les images ou leur circulation (Julliard, 2022) ; enfin, certain·es ont une approche qualitative ou hybride, tandis que d'autres se penchent plus exclusivement sur l'analyse de masses de données (Boyd et Crawford, 2012). Ces pratiques n'échappent pas aux problèmes récurrents de l'ouverture des données – statut public ou privé des échanges en ligne, régimes juridiques applicables, enjeux d'anonymisation ou de consentement (Latzko-Toth et Proulx, 2013). Si des affaires documentées de rupture de confidentialité ont éclaté aux États-Unis (Zimmer, 2010), c'est encore peu le cas en France, ce qui n'empêche pas pour autant le développement d'une réflexion sur ces questions, dans le champ des sciences de l'information et de la communication (Barats dir., 2013), de la sociologie (Beuscart, 2017), des SHS en général (Debaets, 2018), de la sociologie de l'innovation et de l'histoire (Musiani *et al.*, 2019), ou encore de l'archivistique et des bibliothèques (Carlin et Laborderie, 2021), pour ne citer que quelques exemples.

Depuis la loi DADVSI[1] de 2006, en France, l'accès aux archives publiques du web français est uniquement accordé aux chercheur·es accrédité·es auprès des deux organismes habilités, la Bibliothèque nationale de France (BnF) et l'Institut national de l'audiovisuel (INA) – qui ont tous deux créé leur structure d'accompagnement au traitement des données (BnF DataLab et Le Lab à l'INA). Dix ans plus tard, le Règlement général sur la protection des données (RGPD) a accru l'attention vis-à-vis des enjeux de confidentialité des données personnelles. En réponse à un besoin de retours d'expériences, cet entretien croise les points de vue d'une archiviste du web, d'une historienne et d'un ethnographe. Dorothée Benhamou-Suesser (DBS), Fred Pailler (FP) et Valérie Schafer (VS) ont ainsi été interrogés sur les problématiques de l'ouverture des données du web à partir de projets auxquels ils ont collaboré (attentats de 2015, Gilets jaunes, Covid-19, viralité en ligne). Ils reviennent plus particulièrement sur les archives du web qui présentent un cas particulier au regard du web vivant, non seulement en termes de collecte, de sélection et curation, de préservation mais aussi d'accès. La réflexion qui suit s'ancre dans ces développements, depuis des projets de recherche commencés dans les années 2010 jusqu'à la volonté actuelle de penser plus pleinement le *FAIR*[2] *Data* ou encore l'ouverture. En montrant différentes formes d'adaptation progressive et en interrogeant autant la collecte de données que leur analyse ou leur valorisation, cet entretien illustre le fait que la réflexion sur les problématiques de confidentialité et la conscience d'enjeux légaux et éthiques n'ont pas attendu le RGPD pour exister, dans le monde de la recherche comme dans les institutions patrimoniales.

Pour commencer, avec quels types de données travaillez-vous ?

DBS : Nous gérons trois types de données. L'archivage du web par la BnF est une collecte conduite à l'aide de robots logiciels (Illien, 2008) qui copient les pages web et l'ensemble des éléments qui les composent (images, feuilles de style) : ce sont les données collectées à proprement parler. Nous produisons aussi des métadonnées d'ordre technique (données contextuelles comme les requêtes HTTP ou les logs des outils de collecte), ou d'ordre intellectuel et documentaire (thèmes, types d'émetteurs et mots

1. Loi n° 2006-961 relative au droit d'auteur et aux droits voisins dans la société de l'information.

2. Acronyme utilisé pour qualifier des données « *findable, accessible, interoperable, reusable* » (faciles à trouver, accessibles, interopérables, réutilisables). Source : https://fr.wikipedia.org/wiki/Fair_data (consulté le 27/09/2023).

clés utilisés pour décrire les contenus sélectionnés). Enfin, nous travaillons avec des données dérivées, qui sont le résultat d'un traitement ultérieur sur les données collectées (index, statistiques).

VS : Pour l'histoire du web, je travaille avec des données construites qui sont des données statistiques (par exemple, des données d'audience sur les réseaux sociaux numériques quand on étudie la viralité). Elles sont complexes, dans la mesure où on n'a pas toujours accès à la fabrique de ces données, et, par ailleurs, le nombre de *retweets* ou de *likes* sur un réseau social numérique est difficile à interpréter. Ça devient encore plus complexe pour le web archivé : le nombre de *retweets* est figé au moment de la capture du tweet, et on ne peut pas suivre sa postérité. Je travaille aussi avec des métadonnées et des données dérivées fournies par les institutions d'archivage du web (la liste complète des URL) et, bien sûr, avec les contenus web archivés.

FP : L'expression même de « données » est polysémique, suivant le contexte technique ou disciplinaire. Les grandes plateformes numériques ont une conception opérationnelle des données : le terme désigne des contenus numériques (textes, images), les métadonnées qui qualifient ces contenus, ainsi que les personnes qui les partagent. L'agrégation de ces métadonnées est une ressource exploitable commercialement. Mais celles-ci ne racontent pas ce qu'il se passe sur la plateforme de façon générale, ne répondent pas à une question de recherche et ne permettront jamais de le faire directement. La recherche, dès lors qu'elle collecte des données depuis l'API[3] d'une plateforme, travaille toujours un matériau de seconde main, même si ce travail est souvent aussi chronophage que la production *ex nihilo* des données, par exemple, dans une enquête par questionnaire.

Comment les collectes et l'archivage sont-ils organisés en fonction de ces différents types de données, et notamment les données personnelles ?

DBS : La BnF capture le web français à des fréquences variées pour en constituer des échantillons représentatifs, selon deux modalités de collecte. D'une part, des collectes larges annuelles qui portent sur tout le domaine français et s'effectuent à partir d'une compilation de listes de noms de domaines et d'URL. Pour celles-ci, la BnF procède à partir de listes fournies par des bureaux d'enregistrement et des registres comme

3. L'API (*Application Programming Interface*) désigne l'« interface de programmation d'application » qui permet d'avoir accès aux données de la plateforme à laquelle elle est reliée.

l'Association française pour le nommage Internet en coopération (Afnic). Il s'agit de données massives : à titre d'exemple, la collecte large de 2022 a porté sur plusieurs millions de domaines et plusieurs milliards d'URL. Cette volumétrie ne permet pas de qualifier les données personnelles au moment où on les collecte. Il existe, d'autre part, des collectes thématiques ciblées, qui résultent d'un travail de sélection manuelle de contenus. Des expert·es à la BnF ou dans les établissements partenaires, en collaboration sur certains sujets avec des centres d'archives, des associations ou encore des chercheur·es, sélectionnent des contenus à archiver, sur une thématique, avec un objectif de représentativité, en cherchant à capturer les différentes opinions ou modes d'expression présents sur le web à une époque donnée. Par exemple, sur la Covid-19, on s'intéresse à tous les aspects de la pandémie sur le web (point de vue scientifique, politique, l'école à la maison, etc.)[4], déclinés par types d'acteurs, la plupart institutionnels ou associatifs. Sont sélectionnés des sites ou blogs personnels, des comptes sur les réseaux sociaux, notamment ceux de personnalités publiques (écrivain·e, influenceur·se). Des contenus participatifs, avec une expression spontanée d'internautes, sont également collectés, par exemple des *hashtags* Twitter en lien avec l'actualité. Les contenus archivés contiennent ainsi de nombreuses données personnelles qui ne sont pas isolables du reste des données collectées. Les prises de paroles d'anonymes et le caractère participatif du web sont ce qui fait sa singularité par rapport aux médias traditionnels : leur capture participe de l'intérêt des collections de dépôt légal du web pour la recherche. Les sciences participatives, le militantisme en ligne ou encore l'évolution des pratiques d'écriture de soi sur Internet ont retenu très tôt l'attention des sélectionneur·ses, comme en témoigne la publication par la BnF du parcours guidé « (S')écrire en ligne : journaux personnels et littéraires » dans les archives de l'Internet dès 2009[5].

Comment peut-on accéder concrètement aux données du web ?

DBS : Les données ont des statuts distincts au regard de la propriété intellectuelle, et ces statuts conditionnent leur diffusion : depuis 2006, le dépôt légal du web autorise la BnF à collecter le web français sans

4. Voir en ligne le billet « Dans les coulisses de la collecte COVID-19. Entretien sur les pratiques des correspondants du Dlweb » de Véronique Tranchant, Chantal Puech, Sophie Gebeil, Valérie Schafer et Alexandre Faye, publié le 16 novembre 2020 sur le carnet Hypothèses de Web Corpora : https://webcorpora.hypotheses.org/953 (consulté le 01/09/2023).

5. Voir en ligne : https://www.bnf.fr/sites/default/files/2018-11/secrireenligne_parcours.pdf (consulté le 28/08/2023).

autorisation préalable des auteur·es de sites ; mais les contenus collectés à ce titre, données personnelles comprises, sont soumis au droit d'auteur et leur diffusion est restreinte : suivant le Code du patrimoine (article L131-2), ils ne sont consultables que par des chercheur·es accrédité·es, à des fins de recherche scientifique, professionnelle ou personnelle, dans les salles recherche de la BnF et de 26 bibliothèques partenaires[6]. L'accès aux données personnelles contenues dans les archives est donc restreint, et l'obtention et la réutilisation des contenus archivés, par exemple dans le cadre d'une publication, au-delà de l'exception de courte citation, nécessitent l'accord des ayants droit. Les métadonnées documentaires et techniques, ainsi que les données dérivées sont en revanche libres de droit, et peuvent donc être utilisées sans restriction.

FP : Dans le cas d'une étude sur le web vivant, l'accès aux contenus n'est pas le même que dans le cas d'un travail sur les archives du web déjà collectées par le dépôt légal. Il faut des techniques, des méthodes, voire des machines différentes. Suivant les compétences des équipes de recherche et les objectifs de la collecte, on va utiliser le *web scraping* (un robot-logiciel qui automatise la collecte des différents éléments des pages web), ou bien on va copier-coller des captures d'écran, etc. Néanmoins, le *web scraping* est souvent rendu techniquement difficile par les plateformes, parce qu'étant les véritables propriétaires des données, elles protègent délibérément leurs ressources d'un vol éventuel par des services concurrents. Ces quinze dernières années s'est développé le système des API, qui, sous certaines conditions, donne directement accès à la base de données d'une plateforme. X (anciennement Twitter) a ainsi offert ces dernières années un accès spécifique aux chercheur·es pour collecter tweets et métadonnées dans tout l'historique de la base de données, mais toutes les plateformes n'ont pas adopté pour autant ce type de politique.

On évoque souvent le brouillage de la frontière public / privé s'agissant des données du web. Qu'en est-il dans votre pratique ?

DBS : Ce qui est soumis au dépôt légal du web, c'est ce qui est « communiqué au public par voie électronique », publié ou rendu public, à l'exception des contenus diffusés sur des espaces réservés à des publics restreints (intranet, espaces privés des réseaux sociaux numériques par exemple). Le périmètre juridique est clair. D'un point de vue technique, les robots de

6. Voir en ligne : https://www.bnf.fr/fr/selection-partagee-et-acces-en-region-aux-archives-de-linternet/ (consulté le 28/08/2023).

collecte n'accèdent d'ailleurs qu'aux parties publiques du web. D'un point de vue éthique, ce qui rend les choses complexes, c'est que, d'une part, la frontière entre public et privé évolue dans le temps, par exemple certains groupes Facebook de Gilets jaunes[7] étaient publics lorsqu'ils ont été collectés et sont ensuite devenus privés. D'autre part, on peut s'interroger sur la connaissance qu'ont les auteur·es de ces contenus et de notre mission d'archivage. Lorsqu'on laisse des traces numériques sur le web, on n'est pas toujours conscient qu'on est en train de faire un acte de publication, et tous les contenus n'ont pas le même statut. Un blog ou site d'artiste est une publication à part entière, et il y a un enjeu démocratique à pérenniser le compte Facebook d'un candidat·e aux élections. Les enjeux sont différents lorsqu'on collecte des parties de sites ou de réseaux sociaux où tout le monde s'exprime, d'autant que ces derniers brouillent les frontières entre espace public et sphère conversationnelle. La perception de ce qui est partageable en ligne a évolué depuis les premiers temps du web. Le droit de retrait au sens du Code de la propriété intellectuelle (article L. 121-4) peut être invoqué par les personnes concernées. D'ailleurs, aujourd'hui, ce droit de retrait qui existait déjà prend un nouveau sens.

Comment se passe, concrètement, la consultation des archives du web de la BnF ?

DBS : L'accès se fait concrètement *via* une application – appelée « Archives de l'internet » – qui permet de rechercher un site ou une page web à partir de son adresse URL. On peut ensuite choisir une date, visualiser les contenus archivés et naviguer dans les archives comme on le fait sur le web vivant, en cliquant de lien en lien mais dans un contexte daté. En outre, les chercheur·es accueilli·es au BnF DataLab peuvent avoir accès à des corpus de données archivées et des outils permettant de les traiter (Gephi, Hyphe, logiciels d'analyse de texte) à travers un dispositif sécurisé, une sorte de « bulle de consultation », qui empêche la sortie des données. Les données sur lesquelles les chercheur·es peuvent travailler hors de la BnF sont des métadonnées ou des données dérivées libres de droit (Carlin et Laborderie, 2021).
FP : Dans le cadre du projet Buzz-F[8] et de notre collaboration avec le BnF DataLab, Valérie et moi avons utilisé des métadonnées des archives du

7. Voir en ligne le billet « Les Gilets jaunes sous l'œil du dépôt légal numérique » d'Anthony Cerveaux, publié le 15 novembre 2019 sur le carnet Hypothèses de Web Corpora : https://webcorpora.hypotheses.org/750/ (consulté le 28/08/2023).

8. Pour plus d'informations, voir en ligne : https://www.c2dh.uni.lu/fr/projects/buzz-f/ (consulté le 01/09/2023).

web qui avaient été rendues accessibles sur un dépôt GitLab, hébergé par l'infrastructure de recherche IR* Huma-Num. Après s'y être identifié·es, les archivistes partageaient avec nous le code pour l'analyse des données et les jeux de métadonnées. En revanche, les vidéos que nous avons étudiées n'étaient pas partagées sur le dépôt. Il nous a fallu un moment avant de trouver la bonne manière de collaborer à distance et autour des données, sans que celles-ci ne circulent pour autant.

Comment gérez-vous les questionnements éthiques dans votre usage de ces données ?

FP : Se tenir au courant des évolutions juridiques et des bonnes pratiques demande du temps, moins parce que les principes éthiques changeraient constamment que parce que l'accompagnement des projets, le terrain et les méthodes ne cessent eux-mêmes d'évoluer. Les projets qui vont déployer le traitement statistique de gros jeux de données sont peut-être ceux qui correspondent le mieux aux cas prévus par les procédures des comités d'éthique, parce qu'ils traitent des données numériques et relèvent, de fait, du RGPD. Dès lors que l'on adopte une approche ethnographique des pratiques numériques, l'éthique du traitement des données se mêle à l'éthique du terrain qui couvre un champ plus vaste de pratiques, puisqu'elle concerne aussi bien les relations interpersonnelles, le rapport aux réseaux d'informateur·rices, etc. Ici, le bricolage reste de rigueur, notamment pour saisir, chemin faisant, comment la progression sur le terrain et le cadrage légal de la production de données affectent de concert la manière de problématiser et de produire la connaissance.

VS : La question de l'anonymisation se pose pleinement pour moi en 2016 lorsqu'on commence le projet « Archives sauvegarde attentats Paris »[9] et que l'on accède massivement aux données Twitter des réactions des internautes aux attentats de 2015. On peut même accéder à des tweets qui ont été supprimés par leur auteur·e, mais qui ont pu entre-temps être archivés, avant une demande de retrait ou un effacement sur Twitter. Vous imaginez la masse d'informations qu'on peut y trouver qui pose de vraies questions éthiques – par exemple le positionnement des *#jenesuispascharlie*, si l'identité de leurs auteur·es devait être publicisée largement. C'est donc en me retrouvant à travailler sur un sujet extrêmement sensible, celui des réactions aux attentats, que j'entre pleinement dans ces questionnements.

9. Pour plus d'informations, voir en ligne : https://asap.hypotheses.org/168#more-168/ (consulté le 01/09/2023).

Est-ce qu'il y a des règles déjà bien arrêtées ? Aujourd'hui, on est de moins en moins laissés individuellement face à ces questions, ne serait-ce que par le contact qu'on peut avoir avec les institutions d'archivage ou les délégués à la protection des données (DPO) dans les universités. On n'est plus dans une formation autodidacte, même si on l'a été à un moment. On apprend constamment.

FP : Il arrive que la recherche évolue en cours de route, et l'on peut être amené à travailler sur des corpus de données que l'on n'avait pas prévus au départ. Ce supplément de données offre parfois une perspective nouvelle sur l'ensemble du terrain, et permet de mieux saisir, par exemple, les enjeux d'anonymisation spécifiques à ce dernier. Il est alors important de revenir sur les stratégies initiales du projet pour en conserver d'abord l'esprit plutôt que les manières.

Gérez-vous des problèmes d'anonymisation des données ainsi mobilisées ?

DBS : Les chaînes de collecte et de mise à disposition des contenus collectés n'intègrent pas de processus d'anonymisation des données. Celle-ci serait impossible à automatiser compte tenu de la structure et de la volumétrie des données concernées. Les données personnelles ou sensibles peuvent être occasionnellement, et sur demande, isolées de l'ensemble des contenus web collectés et rendues inaccessibles. La question de l'anonymisation se pose surtout à l'étape de leur exploitation par les chercheur·es, au moment d'extraire des corpus restreints hors de la BnF, ou de mobiliser ces données dans une publication. L'indexation plein texte de la totalité des archives, aujourd'hui limitée à quelques corpus, ainsi que la mise à disposition d'outils d'exploration ou de fouille permettant de croiser les différentes données archivées donneront sans doute une importance croissante à ces questions.

FP : Aucun·e chercheur·e ne se positionnera sur la pertinence d'anonymiser un corpus sans interroger d'abord les conditions d'apparition des informations personnelles sur la plateforme ou le service. Somme toute, le plus souvent, il n'y a pas besoin de publier de noms ou d'informations personnelles. Il faut noter néanmoins que les mondes numériques impliquent l'usage délibéré de pseudonymes, et que l'anonymat est toujours contextuel, ce qui complique l'application systématique d'une seule et même règle d'anonymisation : certaines personnes sont plus connues (en ligne ou dans d'autres médias) par leur pseudonyme que par leur nom civil, et il faudra tout autant anonymiser le pseudonyme que l'identité civile,

alors que d'autres personnes vont utiliser un pseudonyme une seule fois et se garantir un anonymat solide en ne le liant jamais à leur identité civile.
VS : J'ai eu le cas sur une recherche sur la question des sites personnels et de l'expression de soi en ligne (Schafer, 2022). Je voulais reproduire une partie des contenus avec l'autorisation des auteur·es. Mais c'est parfois un vrai défi de retrouver les auteur·es de sites ou d'en obtenir une réponse. J'ai dû renoncer à certains contenus et les enlever. Parfois, on bricole pour reproduire une capture, en masquant le nom. Mais l'anonymisation n'est pas systématique. Il y a des moments où elle n'aurait pas de sens. Quand on s'intéresse par exemple aux pratiques des *early adopters* d'Internet et à l'itinéraire dans les années 1990 de quelques figures qui commencent à sortir du lot, à s'investir professionnellement dans le web, si on masque les noms, les parcours, outre le fait qu'on les découvrirait de toute façon, d'un point de vue d'analyse historique, ça n'aurait pas de sens d'anonymiser. Néanmoins, dans d'autres cas la question se pose. Même si on n'est pas capable d'identifier la personne physique, on a une responsabilité dans notre recherche. Pointer un pseudonyme sur Twitter qui peut permettre de remonter tout son fil de discussion, de croiser ses réseaux met cette personne sous le feu des projecteurs ou permet de faire des liens avec d'autres personnes. Même si, nous, on ne sait pas forcément qui est derrière le pseudonyme, notre responsabilité ne disparaît pas, puisque d'autres pourraient le retrouver.

Existe-t-il d'autres moyens pour contrôler la protection des données personnelles ?

FP : Limiter la circulation de contenus n'est pas un problème lié au seul dépôt légal, cela concerne aussi les entreprises qui gèrent les plateformes numériques. Jusqu'à cet hiver et la quasi-fermeture de son API aux universitaires, Twitter suggérait de ne pas publier de corpus complets de tweets et proposait à la place un système de « déshydratation » et de « réhydratation » des corpus. Le principe consistait à laisser les chercheur·es partager des listes comportant les identifiants uniques des tweets, sans pour autant partager les contenus ni l'identifiant de leurs auteur·es. Il fallait donc « déshydrater » les corpus avant de les partager, et Twitter se chargeait d'offrir la « réhydratation » du corpus à d'autres chercheur·es qui se présentaient plus tard avec cette même liste d'identifiants uniques de tweets. Avec cette procédure, Twitter met en scène un intérêt pour l'éthique de la circulation des données, tout en gardant le contrôle sur cette circulation, en s'assurant que les contenus modérés entre-temps ne ressortent pas par

la suite. Ce fonctionnement présuppose que la recherche n'aurait aucune question à poser sur tous ces tweets disparus, ni sur les règles de modération elles-mêmes.

Comment envisagez-vous la mise en œuvre des principes FAIR dans le cadre des restrictions d'accès pesant sur les données archivées ?

VS : Dans le *FAIR Data*, il y a l'idée de reproductibilité. Sur la reproduction de l'analyse, ce qu'on aimerait réussir à faire, c'est, par nos méthodes, donner envie à d'autres de s'emparer des objets qu'on étudie, voire d'affiner nos méthodologies. En pratique, la question de la reproductibilité se pose essentiellement sur de gros traitements et corpus, où on fait une lecture quantitative, et quelqu'un peut vouloir vérifier ou prolonger ce qu'on a fait. Alors, on explique comment a été constitué le corpus. On ne peut pas toujours mettre à disposition toute la liste des URL, encore moins les contenus, mais on définit le type de données utilisées ou le cadre dans lequel elles ont été transmises, ainsi que les outils employés. On n'a pas la possibilité de partager les archives du web et, de plus, on est rarement exclusivement sur ces dernières. Par contre, expliquer comment on a cherché, c'est-à-dire les mots-clés qu'on a utilisés, la sélection, les permaliens vers les ressources, le nombre de résultats qu'on a obtenus, l'espace chronologique dans lequel ils s'inscrivent, c'est déjà une façon de partager un peu.
DBS : Qu'est-il intéressant de partager : les méthodologies de constitution de corpus ou les corpus eux-mêmes ? L'archivage du web permet de pérenniser les corpus et de stabiliser les matériaux qui servent de base aux analyses. Documenter de manière la plus détaillée possible les méthodologies de constitution de corpus permet de rejouer, vérifier, approfondir une analyse.

Finalement, d'où viennent les problèmes liés à l'exploitation de ce type de données ?

VS : Parfois, je reçois des messages où les gens – qui pensent que j'archive le web – me disent : « Je veux savoir si mon site est archivé. » Je pense au cas d'un artiste qui avait tout perdu, et qui s'attendait à ce que tout ait été archivé. J'ai essayé d'aider, en demandant à la BnF, à l'INA s'ils avaient des contenus. Il n'était pas satisfait parce qu'il aurait voulu retrouver l'intégralité de son site. Je crois qu'actuellement la tendance est plus à essayer de demander si on est bien dedans, que de demander à être enlevé. Aux

États-Unis, avec Internet Archive, accessible en ligne, il y a eu quelques affaires, et il y a des pressions pour enlever des contenus. En France, on n'est pas dans le même régime libéral ; tout est encadré par le régime du dépôt légal. Ce genre d'affaires a moins de chances de se produire.
FP : Les problèmes liés à l'exploitation de données se trouvant sur les plateformes ou sur le web vivant sont peut-être bien plus nombreux que ceux liés à l'étude des archives du web. Des cas comme celui de Cambridge Analytica renvoient à des escroqueries intentionnelles et ont reposé sur le fait que les plateformes rechignent rarement à faire commerce des données de leurs usager·ères. Je ne connais pas de jeux de données scientifiques qui aient fuité ou qui aient été hackés en dehors de ce type de configuration. Pour les archives du web, peu de personnes connaissent leur existence et savent en faire usage, il est donc normal que l'on n'ait pas beaucoup d'exemples à raconter. Un usage se développe ces dernières années chez les chercheur·es : il consiste à intégrer dans une étude un petit volet sur les archives du web ou sur les plateformes de sociabilités numériques, en parallèle d'un terrain principal qui porte sur d'autres types de données. Cela marque le fait que le web et les mondes numériques accueillent bien désormais des éléments de la vie quotidienne et que la recherche prend ces éléments en considération. Ce sont ces études secondaires qu'il faut accompagner en ce moment, car, du fait même de leur caractère secondaire, elles pourraient échapper, par manque de temps, à des questions éthiques classiques et à l'accompagnement par les DPO, tout autant qu'à des solutions au cas par cas.

À l'issue de cet entretien collectif, plusieurs points de discussion peuvent être soulevés. D'abord, il entrecroise des retours de chercheur·es et de professionnel·les de la conservation : il est essentiel de prendre en compte toute la chaîne de production et de préservation des données qui forme un écosystème complexe. Il faudrait ainsi pouvoir aussi saisir le rôle des plateformes, de l'édition et des infrastructures de recherche. Ensuite, si les métiers de la recherche sont transformés par l'ouverture des données, ceux des bibliothèques et du patrimoine le sont également en termes de compétences, processus et pratiques, infrastructures, pour ne citer que quelques aspects (Mesguich, 2023). L'entretien insiste également sur le cadre réglementaire, tout en soulignant que celui-ci ne peut pas être toujours adapté tel quel, « clé en main » pour toutes les recherches, et requiert aussi du cas par cas. Anonymiser les données est par exemple pertinent pour une enquête par fouille automatisée de données en SHS, mais ne l'est pas forcément dans une enquête historique sur les premiers utilisateurs du

web. L'entretien insiste ainsi sur la constante tension entre bonnes pratiques et artisanat. Si le terme « bricolage » revient à plusieurs reprises, c'est que, malgré le cadre actuel proposé par le RGPD, les *data management plans* établis au moment des soumissions de projets, les échanges constants avec les DPO, il reste des zones grises, notamment quand il s'agit de publier les résultats, d'organiser leur reproductibilité, ou de partager les données pour une analyse secondaire. Plus récemment, l'écho médiatique rencontré par la fermeture de la plateforme Skyblog, archivée en 2023 par la BnF et l'INA, est le signe d'un intérêt scientifique et public croissant pour ces archives des premiers temps du web ou des premiers réseaux sociaux, qui permettent de retracer l'évolution des pratiques d'écriture de soi et de l'intime, ou de partage de données personnelles. Cet intérêt incite à maintenir un équilibre entre ouverture des données pour les besoins de la recherche et restrictions de consultation dans un souci de protection des données personnelles. Plus généralement, l'entretien souligne les difficultés et les précautions qui doivent être prises pour conserver, diffuser et rendre exploitables des données personnelles dont le statut peut être ambigu. De nombreux dilemmes se posent aux acteur·es ayant à gérer ces données, et les solutions alternatives au dispositif de l'archivage légal (par exemple le *cloud*) ne sont pas sans soulever de sérieuses interrogations. Cet ensemble est aujourd'hui animé par de nombreux espaces de réflexions, à l'image de projets comme ResPaDon[10] en France, ou de l'International Internet Preservation Consortium[11] au-delà, qui témoignent d'une dynamique pour penser les conditions d'accès et d'exploitation de ce type de données.

▮▮▮ références

Barats C. (dir.), 2013. *Manuel d'analyse du web en Sciences Humaines et Sociales,* Paris, Armand Colin, https://doi.org/10.3917/arco.barat.2013.01.

Beuscart J.-S., 2017. Des données du Web pour faire de la sociologie… du Web ?, *in* P.-M. Menger et S. Paye (dir.), *Big data et traçabilité numérique. Les sciences sociales face à la quantification massive des individus*, Paris, Collège de France, 141-161, https://doi.org/10.4000/books.cdf.5009.

10. Réseau de partenaires pour l'exploitation et l'analyse de données numériques. Pour plus d'informations, voir en ligne : https://respadon.hypotheses.org/ (consulté le 01/09/2023).

11. Pour plus d'information, voir en ligne : https://netpreserve.org/ (consulté le 01/09/2023).

Boyd D., Crawford K., 2012. Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon, *Information, Communication & Society*, 15 (5), 662-679, https://doi.org/10.1080/1369118X.2012.678878.

Carlin M., Laborderie A., 2021. Le BnF DataLab, un service aux chercheurs en humanités numériques, *Humanités numériques*, 4, https://doi.org/10.4000/revuehn.2684.

Debaets É., 2018. *Big data* en sciences sociales et protection des données personnelles, *in* V. Ginouvès et I. Gras (dir.), *La Diffusion numérique des données en SHS. Guide des bonnes pratiques éthiques et juridiques*, Aix-en-Provence, Presses Universitaires de Provence, 61-72.

Illien G., 2008. Le dépôt légal de l'internet en pratique : les moissonneurs du web, *Bulletin des bibliothèques de France (BBF)*, 53 (6), 20-27, https://bbf.enssib.fr/consulter/bbf-2008-06-0020-004 (consulté le 09/01/2024).

Julliard V., 2022. Communauté politique, sémiotique, émotionnelle. Ce que la circulation des images révèle de la structuration de la mobilisation anti-genre sur Twitter, *Communication & langages*, 212 (2), 131-153, https://doi.org/10.3917/comla1.212.0131.

Latzko-Toth G., Proulx S., 2013. Enjeux éthiques de la recherche sur le Web, *in* C. Barats (dir.), *Manuel d'analyse du web en Sciences Humaines et Sociales*, Paris, Armand Colin, 32-52, https://doi.org/10.3917/arco.barat.2013.01.0032.

Mesguich V., 2023. *Les Bibliothèques face au monde des données*, Villeurbanne, Presses de l'Enssib.

Musiani F., Paloque-Bergès C., Schafer V., Thierry B. G., 2019. *Qu'est-ce qu'une archive du web ?*, Marseille, OpenEdition Press, https://doi.org/10.4000/books.oep.8713.

Schafer V., 2022. Préserve-moi ! Des journaux intimes à ceux de confinement dans les archives du Web, *Le Temps des médias*, 38 (1), 175-194, https://doi.org/10.3917/tdm.038.0175.

Zimmer M., 2010. "But the Data is Already Public": on the Ethics of Research in Facebook, *Ethics and Information Technology*, 12 (4), 313-325, https://doi.org/10.1007/s10676-010-9227-5.

Dorothée Benhamou-Suesser est chargée de collections numériques au service du dépôt légal numérique de la Bibliothèque nationale de France (BnF), responsable de la préservation et des outils d'accès aux Archives de l'internet. À ce titre, elle travaille en étroite relation avec les équipes du département des systèmes d'information de la BnF ainsi qu'avec les chercheurs qui utilisent ces outils.

▮ dorothee.benhamou-suesser@bnf.fr

Fred Pailler est chercheur postdoctoral dans le cadre du projet HIVI, au Center for Contemporary and Digital History (C^2DH). Titulaire d'un doctorat en sciences de l'information et de la communication (SIC), il est spécialiste d'enquêtes combinant approche ethnographique, questionnaires et analyse de données numériques. Ses recherches portent sur les controverses liées au travail, à la santé, au genre ou aux sexualités, notamment sur les grandes plateformes sociales.

▮ frederic.pailler@uni.lu

Valérie Schafer est professeure d'histoire européenne contemporaine au Center for Contemporary and Digital History (C^2DH) à l'université du Luxembourg. Spécialiste de l'histoire du numérique et des archives du web, elle a coordonné de nombreux projets s'appuyant sur les données du web (Web90, ASAP sur les attentats de 2015, AWAC2 sur la Covid-19, HIVI sur l'histoire de la viralité en ligne). Elle est également membre du comité scientifique du Lab de l'INA.

▮ valerie.schafer@uni.lu

Selma Bendjaballah est docteure en science politique et ingénieure de recherche au Centre de données socio-politiques (CDSP) de Sciences Po. Elle assure plusieurs enseignements de science politique et de méthodologie qualitative. Après s'être consacrée à l'enrichissement et à la valorisation de la Banque d'enquêtes qualitatives en sciences humaines et sociales (beQuali) du CDSP, elle s'occupe des activités de valorisation et de recherche méthodologiques.

▮ selma.bendjaballah@sciencespo.fr

Guillaume Garcia est docteur en science politique et ingénieur de recherche au Centre de données socio-politiques (CDSP) de Sciences Po. Après avoir géré le développement de la Banque d'enquêtes qualitatives en sciences humaines et sociales (beQuali) du CDSP, il s'occupe des activités de recherche méthodologique et de valorisation au sein du laboratoire. Ses travaux actuels portent sur les enjeux de mise à disposition de corpus d'enquêtes qualitatives à des fins de recherche ou d'enseignement.

▮ guillaume.garcia@sciencespo.fr