Inflation des données, renégociations des cadres, opportunités pour la recherche en SHS ?

Ghislaine Chartron

questions de communication

Questions de communication

41 | 2022
20 ans, 10 questions, 20 réponses

Inflation des données, renégociations des cadres, opportunités pour la recherche en SHS ?

Data inflation, framework renegotiations, opportunities for Social Sciences and Humanities research?

Ghislaine Chartron

OpenEdition Journals

Édition électronique
URL : https://journals.openedition.org/questionsdecommunication/28334
DOI : 10.4000/questionsdecommunication.28334
ISSN : 2259-8901

Éditeur
Presses universitaires de Lorraine

Édition imprimée
Date de publication : 1 juin 2022
Pagination : 391-406
ISBN : 978-2-38451-018-4
ISSN : 1633-5961

Référence électronique
Ghislaine Chartron, « Inflation des données, renégociations des cadres, opportunités pour la recherche en SHS ? », *Questions de communication* [En ligne], 41 | 2022, mis en ligne le 01 octobre 2022, consulté le 25 octobre 2022. URL : http://journals.openedition.org/questionsdecommunication/28334 ; DOI : https://doi.org/10.4000/questionsdecommunication.28334

questions de communication, 2022, 41, 391-406

GHISLAINE CHARTRON
Conservatoire national des arts et métiers, Dicen-IDF, F-75003 Paris, France
ghislaine.chartron@lecnam.net

INFLATION DES DONNÉES, RENÉGOCIATIONS DES CADRES, OPPORTUNITÉS POUR LA RECHERCHE EN SHS ?

Résumé. — Cet article s'intéresse au développement des données numériques et aux évolutions potentielles pour les recherches en sciences humaines et sociales. Il dresse des éléments du renouvellement du cadre technique et juridique pour les terrains mobilisant différents types de données. L'enjeu de la mise à disposition de services simplifiés et intégrés aux environnements numériques est identifié comme une condition d'acculturation à cette nouvelle ouverture. Il relève quelques usages significatifs de fouilles de données en sciences humaines et sociales mais mesure également les limites de ces approches quantitatives. Les questions épistémologiques sous-jacentes sont rappelées et l'auteure plaide pour une hybridité de méthodes, les données massives ouvrant une nouvelle voie d'analyse du social avec ses enjeux et ses limites.

Mots clés. — fouille de données, pratiques numériques, cadre technique, cadre juridique, épistémologie, science des données, sciences humaines et sociales

L'environnement de la recherche a changé. Les humanités numériques se sont installées, les données se sont multipliées, les méthodes numériques font désormais l'objet de manuels dédiés (Van Hooland *et al.*, 2016). Le croisement des sciences humaines et sociales (SHS) et du numérique n'est toutefois pas nouveau et s'inscrit dans l'histoire de l'évolution des outils. Certaines disciplines ont été pionnières comme l'archéologie, les études médiévales, la géographie, les sciences de l'information. Qu'est-ce qui a changé depuis ces dernières années ? On assiste à un saut considérable que l'on pourrait expliquer par la convergence de plusieurs dynamiques : l'inflation des données produites par les individus, les organisations, les sociétés (traces de consultation du Web, interactions en ligne, *open data*, objets connectés…), une croyance politique affirmée dans la « société des données » et ses bienfaits sur l'économie, l'innovation[1], la démocratie et enfin l'appropriation progressive d'outils pour « faire parler » les données, leur donner du sens et anticiper l'avenir par un apprentissage sur les masses de données déjà disponibles.

Dans cette contribution, nous voudrions apprécier objectivement les avancées et les opportunités du numérique pour le travail du chercheur en SHS, mais pointer également les limites de telles approches. Les extrêmes sont toujours un peu aveugles. Il s'agit de défendre une vision équilibrée de cette hybridation renouvelée avec le numérique. Nous nous distinguerons donc d'une posture exclusivement critique qui tend à voir dans les outils numériques des prismes déformants de la compréhension du social (Ouellet, 2021) et nous n'emprunterons pas non plus la voie du prosélytisme des ingénieurs ou des politiques dont la croyance en la technologie les éloigne trop souvent des terrains, de leurs besoins et de leurs priorités.

Nous insisterons sur la dimension « donnée numérique » devenue dominante et transversale aux champs scientifiques. Nous pointerons, dans un premier temps, le renouveau des sources. Puis, il s'agira de souligner l'évolution du cadre technique et du cadre juridique pour la fouille de données. Nous terminerons par des questions épistémologiques fondamentales qui traversent aujourd'hui les sciences humaines et sociales : primauté de la théorie, pertinence des échantillons, corrélation ou causalité…

Renouveau des sources de données

Le renouveau des sources de données conjugue plusieurs mouvements : les politiques publiques de soutien aux données, la centralité de nos activités sur le Web et les multiples traces associées, la transformation digitale des organisations

1 L'innovation est au cœur de la stratégie européenne pour les données définie par la Commission européenne en février 2020. Accès : https://ec.europa.eu/info/sites/default/files/communication-european-strategy-data-19feb2020_fr.pdf (consulté le 19 janv. 2022).

et leur production de données, enfin l'insertion de capteurs et d'objets connectés dans nos vies quotidiennes.

Dans le contexte de la société numérique, des promesses de l'intelligence artificielle (IA) et de ses innovations, des enjeux citoyens de la transparence publique, les politiques de la donnée se sont déployées, en particulier pour les données produites par les administrations (*open data*) et les données de la recherche pour ce qui nous intéresse. La politique de la Commission européenne s'est ainsi traduite par plusieurs recommandations et évolutions du cadre règlementaire que les États membres ont dû transposer. Les garde-fous ont été également élaborés concernant la protection de la vie privée (règlement général sur la protection des données, RGPD) et l'encadrement de l'IA au regard des valeurs humaines fondamentales. Une approche par le risque est ainsi inscrite dans le règlement IA en cours d'adoption[2]. Concernant l'*open data*, partie intégrante de l'*Open Government Initiative*[3] lancé en 2009 par l'administration Obama, la France diffuse aujourd'hui plus de 40 000 jeux de données (voir figure 1). Le gouvernement a programmé une nouvelle accélération avec la directive Castex d'avril 2021[4], demandant aux ministres de *« [s'impliquer] personnellement »* pour faire de la politique de la donnée *« une priorité stratégique de l'État »*. On peut donc s'attendre à disposer de jeux de données encore plus nombreux même si la quantité ne dit rien sur la valeur et la réutilisation effective de ces données. Divers baromètres s'attachent à mesurer cette progression et à comparer les pays pour promouvoir le mouvement[5].

Figure 1. Quantification de l'offre Open data française (janvier 2022, accès : https://www.data.gouv.fr/fr/)

L'*open science* est un autre mouvement mondial et conjoncturel qui se nourrit également d'une vision d'efficience et d'intégrité de la science, vision relayée et promue en France par le Comité pour la science ouverte[6] (Coso). Les publications scientifiques et les données de la recherche financées sur fonds publics doivent

2 La proposition de règlement du Parlement européen et du conseil établissant des règles harmonisées concernant l'intelligence artificielle a été établie le 21 avr. 2021. Accès : https://eur-lex.europa.eu/legal-content/FR/ALL/?uri=CELEX:52021PC0206 (consulté le 19 janv. 2022).

3 Accès : https://fr.wikipedia.org/wiki/Open_Government_Initiative (consulté le 19 janv. 2022).

4 La directive Castex correspond à la circulaire n°6264/SG du 27 avr. 2021 relative à la politique publique de la donnée, des algorithmes et des codes sources. Accès : https://www.legifrance.gouv.fr/circulaire/id/45162 (consulté le 19 janv. 2022).

5 Accès : https://www.etalab.gouv.fr/retour-sur-lopen-data-maturity-index-2021-politique-gouvernance-de-lopen-data-en-france-1-4 (consulté le 19 janv. 2022).

6 Accès : https://www.ouvrirlascience.fr (consulté le 19 janv. 2022).

être accessibles sans barrière. HAL diffuse ainsi en accès ouvert plus d'un million de documents scientifiques début 2022. Là encore, les baromètres se multiplient, celui du Coso veut mesurer tous les ans les progrès de l'accès ouvert en France. Le Coso précise ainsi : « Selon l'édition 2021 du baromètre français de la science ouverte, 62 % des 166 000 publications scientifiques françaises [suivant les affiliations détectées des auteurs] publiées en 2020 sont en accès ouvert en décembre 2021[7] ».

La centralité du Web dans nos activités citoyennes, professionnelles, de loisir et de consommation en fait également aujourd'hui une source de données majeure (logs de navigations, d'usages de ressources, de commentaires et d'avis…). Les données publiques des réseaux sociaux sont particulièrement prisées (Twitter notamment). Les données du Web concernent aussi toutes les données de contenu des sites Web en accès public et sous licence. Quant à la transformation digitale des entreprises, elle s'accompagne d'une production de données internes que l'on cherche de plus en plus à valoriser dans une finalité de performance (améliorer la relation client, les activités métiers, la productivité…) (Chartron, 2021). Les tableaux de bord générés veulent assister de plus en plus le pilotage et l'aide à la décision. Enfin, les données des capteurs et des objets connectés notamment dans le secteur de l'e-santé (santé numérique) et du *quantified self* sont aussi de nouvelles sources qui peuvent intéresser le chercheur en anthropologie ou en sociologie, et le téléphone portable, avec toutes ses applications dédiées, est devenu le capteur central.

Ce bref panorama, certainement incomplet, donne un aperçu du renouveau des sources de données. La mise à disposition n'implique cependant pas une valeur d'usage (Rebouillat, 2021) comme nous le verrons plus loin.

Le cadre technique des « terrains données »

Les étapes des « terrains données » sont assez communes : collecte, nettoyage, traitement par des analyses statistiques, restitution des résultats sous forme de visualisations graphiques. Nous allons brièvement les résumer en insistant sur certaines dimensions renouvelées.

Les techniques de collecte de données se sont diversifiées en fonction des différentes sources que nous avons évoquées précédemment : téléchargement de fichiers dans des formats dominants (format CSV, format JSON) notamment pour l'*open data*, transfert de données par des API (interfaces de programmation) comme pour les données Twitter[8], requêtage en langage SQL pour les bases de

7 Accès : https://www.ouvrirlascience.fr/le-barometre-de-la-science-ouverte/ (consulté le 10 mai 2022).

8 Accès : https://developer.twitter.com/en/docs (consulté le 19 janv. 2022).

données comme pour les systèmes d'information des organisations. Le *Web scraping* (le « grattage du Web ») renouvelle aussi cette phase de collecte de données : la technique concerne l'extraction de données de sites Web par des algorithmes adaptés, *via* le protocole de transfert hypertexte (HTTP) ou par le biais de navigateurs Web. On peut ainsi extraire des adresses, des liens, des articles scientifiques… Le Web offre une panoplie de données librement accessibles par des tiers.

Le nettoyage des données est l'étape suivante, essentielle pour pouvoir explorer ensuite les données, en extraire du sens selon des méthodes statistiques. Les opérations concernent par exemple le dédoublonnage, les corrections de données erronées ou manquantes, la normalisation, l'intégration de données différentes… puis le formatage pour le *data mining*.

Selon la définition de la nouvelle directive européenne sur le droit d'auteur et les droits voisins, le *data mining* (ou fouille de données) désigne « toute technique d'analyse automatisée visant à analyser des textes et des données sous une forme numérique afin d'en dégager des informations, ce qui comprend, à titre non exhaustif, des constantes, des tendances et des corrélations[9] ». La pratique est déjà ancienne si l'on se réfère à l'école française d'analyse des données de Jean-Paul Benzécri des années 1960 (Beaudouin, 2016) et aux approches d'analyse textuelle des corpus par lexicométrie des années 1980 (Salem, 1982). Où se situe donc la nouveauté ? Une offre logicielle plus accessible à des non-programmeurs (comme les logiciels Iramuteq, Tableau, Qlik Sense ou Power BI par exemple) a rendu plus abordables ces méthodes d'exploration quantitatives. La masse des données aujourd'hui disponible permet également la mobilisation de techniques d'apprentissage à des fins prédictives.

La fouille de données est étroitement liée à l'étape finale de datavisualisation, transmission des résultats de manière attrayante à l'aide de techniques de visualisation mobilisant les fondamentaux de la sémiotique visuelle. La datavisualisation se trouve également popularisée aujourd'hui en grande partie par la disponibilité d'une offre logicielle abondante et de plus en plus intégrée aux logiciels de fouille de données.

Un cadre juridique qui devient facilitateur pour les chercheurs

Le risque juridique a souvent freiné les motivations des chercheurs pour investir « un terrain données » dans leur travail scientifique. Aujourd'hui, les

9 Accès : https://eur-lex.europa.eu/legal-content/FR/TXT/PDF/?uri=CELEX:32019L0790&from=FR (consulté le 19 janv. 2022).

conditions d'usage se sont considérablement ouvertes à la suite de l'évolution des encadrements juridiques qu'il semble important de rappeler.

D'abord, l'encadrement par des licences ouvertes attachées à la majorité des données publiques et textes scientifiques de revues en *open access*. En 2017, on dénombrait 1,4 milliard d'œuvres sous licence *Creative Commons*[10] au niveau mondial. La licence CC0 permet aux producteurs de placer leurs données dans le domaine public, sans aucune restriction d'usage. Chaque État a également élaboré des licences particulières. En France, la loi « pour une République numérique » de 2016 a énuméré une liste de licences utilisables par les administrations pour diffuser leurs données[11]. La licence d'Etalab est considérée comme une licence de référence pour la publication des données publiques.

Par ailleurs, une avancée majeure concerne la nouvelle exception au droit d'auteur dans le cadre de la directive européenne du droit d'auteur de 2019[12] que la France a transposé le 24 novembre 2021[13], autorisant désormais, pour les chercheurs, l'exploration automatisée de textes et de données (*text and data mining* ou TDM) protégés par des droits de propriété intellectuelle. C'est une nouvelle exception au droit d'auteur en faveur de la recherche et qui permet au chercheur de conduire des fouilles de données sans avoir besoin d'obtenir l'autorisation des producteurs des textes. La condition à respecter est d'accéder à ces données avec un accès licite lorsque ces données sont payantes, l'abonnement des universités peut le permettre. Les titulaires de droits sur les données sont par ailleurs autorisés à mettre en œuvre « des mesures proportionnées et nécessaires afin d'assurer la sécurité et l'intégrité des réseaux et des bases de données dans lesquels les œuvres sont hébergées » (texte de la loi). Il peut s'agir de conditionner l'accès par des API ou de limiter la quantité de données par transaction. Nous renvoyons au billet de Lionel Maurel et Stéphanie Rennes pour des précisions complémentaires[14].

Concernant le *Web scraping* de sites en libre accès pour des tiers, les règles juridiques sont plus floues car plusieurs cadres juridiques sont mobilisables : le droit d'auteur mais aussi la directive du Parlement européen et du Conseil du 11 mars 1996 sur la protection juridique des bases de données et sa transposition

10 Accès : https://stateof.creativecommons.org/ (consulté le 19 janv. 2022).

11 Accès : https://www.data.gouv.fr/fr/pages/legal/licences/ (consulté le 19 janv. 2022).

12 Accès : https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=celex%3A32019L0790 (consulté le 19 janv. 2022).

13 Accès : https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000044362034 (consulté le 19 janv. 2022).

14 L. Maurel et S. Rennes, « La fouille de textes et de données à des fins de recherche : une pratique confirmée et désormais opérationnelle en droit français », billet, *Ouvrir la science !*, 16 déc. 2021. Accès : https://www.ouvrirlascience.fr/la-fouille-de-textes-et-de-donnees-a-des-fins-de-recherche-une-pratique-confirmee-et-desormais-operationnelle-en-droit-francais/ (consulté le 19 janv. 2022).

française[15]. La recommandation de la Commission nationale de l'informatique et des libertés (CNIL) est de recueillir l'autorisation du producteur comme dans le cas du RGPD[16]. Le *Web scraping* peut dégrader considérablement les performances techniques des sites Web. De nombreux producteurs ont spécifié dans leurs conditions générales d'utilisation (CGU) les modalités associées à des collectes automatisées de leurs données, c'est par exemple le cas de Facebook qui interdit la collecte automatisée sans un accord encadré[17]. Twitter, contrairement à Facebook, promeut la réutilisation de ses messages publics (et non privés) en mettant à disposition des API spécifiques pour les réutilisateurs[18]. Concernant le *scraping* qui serait fait dans un cadre non académique, l'article 4 de la nouvelle directive du droit d'auteur laisse la possibilité au producteur de poser des limites à la collecte automatisée. Un groupe de travail est en cours au niveau du W3C (World Wide Web Consortium) pour permettre une lecture automatique de ces consentements par les algorithmes scrapeurs[19], l'objectif étant la mise en place d'un standard *Legal Tech* au niveau du W3C. La loi serait ainsi encryptée dans le code selon le principe « *Code is Law*[20] ».

Pour terminer ce volet juridique, il faut rappeler le cadre général pour traiter des données personnelles (Directive européenne RGPD de 2018, transposée en France en juin 2019). Le consentement explicite des personnes est à recueillir ainsi que la diffusion d'une mention précisant le responsable du traitement, la finalité, la nature des données collectées, le respect de la confidentialité et les modalités de conservation et sécurisation des données. Le lecteur trouvera des précisions dans le guide élaboré par l'Institut des sciences humaines et sociales (InSHS) du Centre national de la recherche scientifique (CNRS) (InSHS, 2021).

---

15 Accès : https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=celex%3A31996L0009 (directive, consultée le 19 janv. 2022) ; https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000000573438 (transposition, consultée le 19 janv. 2022).

16 PLR Avocats, « Warning : Web scraping et RGPD », 14 mai 2021. Accès : https://www.plravocats.fr/blog/data-protection-rgpd/warning-web-scraping-et-rgpd (consulté le 19 janv. 2022).

17 Facebook, *Automated Data Collection Terms* (Conditions de collecte de données automatisées), 15 avr. 2010 (date de la dernière révision). Accès : https://www.facebook.com/apps/site_scraping_tos_terms.php (consulté le 19 janv. 2022) ; Proxy VPN, « Est-il possible de scraper des datas sur Facebook et comment ? ». Accès : https://www.proxyvpn.fr/scraper-facebook (consulté le 19 janv. 2022).

18 Twitter, Centre d'assistance, « À propos des API Twitter ». Accès : https://help.twitter.com/fr/rules-and-policies/twitter-api (consulté le 19 janv. 2022) ; Twitter, Developer Platform, « Tap into what's happening to build what's next ». Accès : https://developer.twitter.com/en (consulté le 19 janv. 2022).

19 Accès : https://www.w3.org/community/tdmrep/ (consulté le 19 janv. 2022).

20 Nous remercions Thomas Saint-Aubin de la société Seraphin.legal et Jean-Baptiste de Vathaire de Cairn.info pour leurs éclairages sur ce sujet.

L'enjeu de l'innovation servicielle pour l'acculturation aux données

Les chercheurs en SHS ont besoin que se développe une offre de services tenant compte de leurs demandes et leur permettant de s'acculturer à la fouille de données, en mesurer les apports et les limites pour leurs recherches. L'émancipation du code et des informaticiens, même si elle ne sera pas toujours possible, doit rester un horizon à atteindre pour croiser réellement le travail scientifique à ces nouvelles opportunités. L'innovation servicielle doit être au rendez-vous.

En analysant les missions de plusieurs promotions d'apprentis en formation de *data analyst*, nous avions pu constater que l'autonomie des métiers face à la valorisation de leurs données était encore embryonnaire mais s'amorçait notamment par la mise à disposition de logiciels dédiés tels que Tableau, Qlik Sense, Power Bi capables de se connecter à différents formats de données et de proposer des fonctionnalités d'analyse statistique et de datavisualisation (Chartron, 2021).

Concernant les chercheurs en SHS, les infrastructures publiques développées dans le cadre du soutien aux humanités numériques, telles que Huma-Num[21] en France, proposent un ensemble de services sur les données. Au niveau européen, le service EOSC (*European Open Science Cloud*[22]) a également pour ambition de mutualiser des ressources pour les communautés scientifiques même s'il est, pour le moment, un peu difficile d'y démêler l'ensemble. Par ailleurs, au niveau des programmeurs, la même tendance simplificatrice s'est affirmée par le recours à des bibliothèques de langage R (comme Shiny) ou de langage Python, en *open source* et dédiées à la fouille de données.

Une fonctionnalité attendue pour les recherches en SHS concerne la possibilité d'explorer des corpus de textes sélectionnés par les chercheurs et pouvant enrichir la compréhension d'une problématique donnée notamment en suivant l'évolution de certains termes, la détection automatique de thématiques couvertes dans un corpus. Ces outils ne sont pas encore « greffés » facilement dans l'environnement de travail. Le service Istex[23], par exemple, permet de télécharger un ensemble de textes dans le périmètre couvert, mais il n'offre pas d'outils d'analyse. Son partenariat avec la plateforme Gargantex est prometteur pour tendre vers cette intégration dans la proposition du service Easistex[24] :

21 Accès : https://www.huma-num.fr/ (consulté le 19 janv. 2022).

22 Accès : https://eosc-portal.eu/about-eosc-portal (consulté le 19 janv. 2022).

23 Accès : https://www.istex.fr/ (consulté le 19 janv. 2022).

24 Accès : https://www.istex.fr/easistex/ (consulté le 19 janv. 2022).

> « [Gargantext] combine des outils issus du traitement du langage naturel, de l'exploration de texte, de l'analyse des réseaux complexes et de la visualisation interactive de données. Il propose donc de nouveaux types d'interactions avec les grands corpus numériques (documents scientifiques, articles de presse, blogs, etc.). [Il] permet de faire, en quelques minutes, des cartes de la connaissance, des états de l'art collaboratifs […] [et peut se connecter] à plusieurs fournisseurs de données *via* des API » (*ibid.*).

Un autre service innovant dans ce registre est le service *Constellate* pour des corpus anglophones, il a été initié par l'organisation JSTOR/ITHAKA afin d'outiller les chercheurs en SHS pour des besoins en fouille de données. La figure 2 montre une datavisualisation construite en dynamique sur la requête « *Datamining* », elle identifie les domaines mobilisant cette technique au regard du corpus couvert (ici, il s'agit du corpus de JSTOR et Portico principalement). On y constate que le sujet concerne majoritairement les sciences appliquées, les sciences biologiques, les mathématiques appliquées, les sciences médicales, l'astronomie… Et pour les SHS, il s'agit des sciences de l'information en premier, du commerce, du droit, de la psychologie, de la communication, de la philosophie, des sciences de l'éducation.

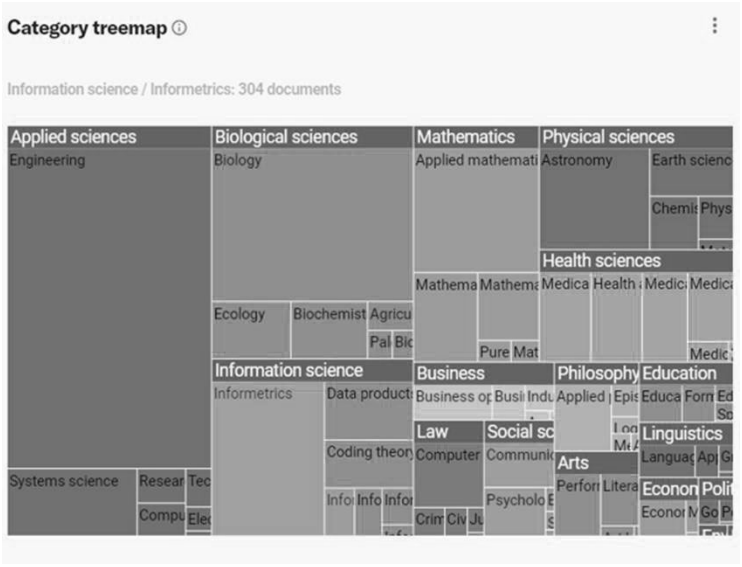

Figure 2. Datavisualisation extraite de la réponse à la requête « Datamining » posée sur le service Constellate de JSTOR/Ithaka

Les pratiques en SHS : quelques tendances

L'objectif n'est pas ici de rendre compte par une méta-analyse de l'ensemble des pratiques, des appropriations de méthodes statistiques et algorithmiques

dans le travail des chercheurs en SHS, ce qui nécessiterait un travail important d'investigation. Nous soulignerons quelques tendances au regard des travaux publiés dans la littérature scientifique et par une observation participante à certains cercles d'échanges.

Face à la croissance des services numériques et des interactions en ligne, les études de publics ont nécessité la prise en compte de la nouvelle donne des traces numériques et en particulier l'analyse des logs des utilisateurs pour comprendre leur comportement. Deux exemples significatifs concernent le champ des sciences de l'éducation et celui des sciences de l'information et de la communication (SIC). Les sciences de l'éducation ont mobilisé la fouille de données pour analyser le comportement des apprenants dans les environnements numériques d'apprentissage en ligne, c'est le champ des *learning analytics* (Baek et Doleck, 2021). De la même façon, les SIC ont mobilisé de telles analyses pour comprendre l'usage des bibliothèques numériques et affiner les services (Chevallier, 2018). Ces méthodes ne se sont pas substituées aux approches qualitatives de type enquêtes mais ont été souvent considérées comme complémentaires, l'intérêt se situant au niveau de la masse des données recueillies par rapport à des enquêtes fondées sur des échantillons et des déclarations subjectives. Une autre application majeure concerne les études bibliométriques qui, par essence, s'intéressent aux mesures quantitatives de la production scientifique.

Un constat concerne également l'attention portée désormais à la fouille des données des médias sociaux (Twitter, Facebook, autres réseaux sociaux, blogs, plateformes de débats…), particulièrement en communication et en sciences politiques (Longhi, 2020), phénomène conjoncturel à ces nouvelles agoras. De nouvelles méthodes de travail se déploient en journalisme (Joux et Bassoni, 2018), le *computational Journalism* donne naissance à de nouvelles équipes comme le *Computational Journalism Lab* de Northwestern University[25].

De façon transversale aux différents champs des SHS, on peut avancer que la mobilisation de la fouille de textes connait au moins trois applications majeures : l'analyse de réseaux sémantiques qui permet d'apprécier les termes les plus structurants et leurs relations, l'identification des thèmes d'un corpus de textes par des algorithmes de classification, et l'analyse des sentiments qui vise à déterminer automatiquement si le sentiment dégagé par une phrase est positif ou négatif. Certaines publications ont insisté sur l'intérêt de telles méthodes pour étudier les « dynamiques humaines » à plusieurs niveaux (Zhang *et al.*, 2020) : individuel, relationnel et collectif. Cependant, dans le contexte de la recherche française, le décomptage actuel sur HAL montre que ce type d'approche reste peu mobilisé par la communauté française en SHS puisque la recherche avec les mots clés « fouille de données » « *data mining* ou *datamining* », « fouille de

25 Le *Computational Journalism Lab* est dirigé par N. Diakopoulos. Accès : https://cj-lab.org/ (consulté le 19 janv. 2022).

textes », « *text mining* » identifie seulement 263 publications. HAL héberge au moment de cette recherche (réalisée le 10 janvier 2022) 281 133 publications en SHS, le ratio est donc de 0,09 %… De façon comparative, mais sans pouvoir le mesurer précisément, le terrain de l'entreprise semble mobiliser davantage ce type de méthodes pour ses données internes, en phase avec la mise en place de tableaux de bord dynamiques pour des objectifs de pilotage, de connaissance client, de performance, de détection des fraudes, d'anticipation des risques (Chartron, 2021)… La disponibilité d'une offre logicielle dédiée a, en grande partie, permis ces développements et les GAFAM y règnent pleinement (Google Analytics, Amazon Web Services, Microsoft Power BI).

La question des données de recherche et de leur réutilisation mérite également une attention particulière. Depuis ces dernières années, on assiste à un discours politique très prescriptif en matière de « données de la recherche » : il faut les sauvegarder, les documenter pour permettre leur réutilisation. Le discours est trop souvent sans nuance, sans questionnement fondamental sur ce qu'est une donnée de la recherche, sur le potentiel réel de la réutilisation. Nous avions décrypté de façon critique cette vision qui pouvait conduire à une bureaucratisation supplémentaire pouvant engendrer réticences et coûts inflationnistes. Nous plaidions pour substituer à une politique descendante une politique d'auto-détermination des communautés scientifiques (Chartron, 2018). De nombreux travaux se sont attachés également à montrer que le concept de « données de la recherche » était difficile à cerner, très hétérogène selon les disciplines (Cabrera, 2014), voire impossible à définir dans l'absolu car très lié à chaque contexte de production (Borgman, 2015) et à l'appréciation du chercheur (Leonelli, 2015). En SIC, Hélène Prost et Joachim Schöpfel (2019) ont exposé cette même difficulté à identifier des archives de données dans ce domaine. Toutefois, on peut distinguer trois macro-catégories : les données externes produites par d'autres chercheurs, les données sources collectées par le chercheur, les données de résultats produites par le chercheur.

Les faibles pratiques de dépôt des chercheurs ont été approfondies par des études qualitatives, mettant en évidence la dimension essentielle de reconnaissance symbolique du travail du chercheur et de la nécessité de laisser du temps d'exclusivité pour l'exploitation des données collectées et construites (Rebouillat, 2021). De façon paradoxale, il est difficile de recueillir aujourd'hui des données sur la réutilisation des jeux de données stockés dans les infrastructures dédiées telles que Progedo en France. Les microterrains nous en disent plus. Ainsi, l'enquête menée au sein du Laboratoire d'économie et de sociologie du travail (Bonneville *et al.*, 2021) montre que les chercheurs en économie, sociologie, sciences politiques, anthropologie utilisent des méthodes qualitatives et quantitatives, que l'économie mobilise le plus les méthodes quantitatives, que trois quarts des chercheurs déclarent ne jamais avoir utilisé des données collectées par d'autres collègues, que les principales sources publiques sont les grandes enquêtes nationales accessibles depuis la plateforme Progedo, ou

d'autres enquêtes d'organismes publics (ministères, OCDE, Eurostat, Céreq…). Ces chercheurs utilisent également des données privées.

Ces constats montrent que la sélection des données reste essentielle, centrée sur l'appréciation du chercheur où confiance et qualité des données sont prioritaires. Des travaux ont approfondi plus en détail les facteurs influençant la réutilisation des données en sciences sociales (Gonçalves Curty, 2015).

Un renouveau méthodologique et épistémologique pour les SHS ?

S'il est aisé de comprendre les enjeux et les opportunités de capitalisation des données dans les sciences de la nature (données collectées par des capteurs ou autres instruments de mesure, données interprétées selon des protocoles communs), il est plus difficile d'envisager la même transposition pour les sciences sociales. La collecte des données d'enquête par exemple est liée à des variables choisies, contextuelles à l'orientation de la recherche : les données sont difficiles à réutiliser. Les corpus de textes peuvent probablement mieux s'aligner sur cet objectif de capitalisation même si des biais sont toujours identifiables. La diversité des données disponibles peut toutefois, selon le type de recherche et la nature des données, ouvrir des possibilités aux chercheurs pour faciliter, consolider, amplifier leurs résultats scientifiques.

Ce renouveau s'accompagne-t-il également d'une remise en cause des méthodes de collecte des données ? Autrement dit, les échantillons, les sondages ont-ils encore un sens dans un contexte où de grands ensembles d'individus peuvent être observés ? Les échantillons et les sondages sont certainement des artefacts imparfaits, marqués par la subjectivité de leurs concepteurs. Mais les données massives ont aussi leurs biais (contexte de captation, données de contribution non représentatives d'une population, données manquantes, hétérogènes…) et leur nettoyage est d'ailleurs la tâche la plus chronophage avant de vouloir les « faire parler » (Berti-Équille, 2018). Par ailleurs, des données massives ne sont pas toujours mobilisables, elles se concentrent sur le Web (logs, commentaires, prises de parole…) et le Web n'est pas le terrain unique d'observation. Les chercheurs en sciences sociales savent également que les méthodes qualitatives sont les plus riches, indispensables pour comprendre des phénomènes sociaux. Le social est complexe, avec des contradictions, des paradoxes, des croyances, des représentations que ne peuvent recueillir les approches quantitatives et rationnelles (Kitchin, 2014). Pour les sciences humaines et sociales, c'est donc bien dans la complémentarité des méthodes que le chercheur trouvera certainement un renouveau. Les données massives permettent de comprendre à un niveau macro certaines tendances majeures, voire d'orienter ses hypothèses. C'est une nouvelle offre d'interprétation du social à considérer mais elle ne peut suffire.

L'échelon supérieur, très controversé, est celui de la potentielle remise en cause de l'épistémologie du travail scientifique généralement organisé par une démarche déductive en sciences sociales, les étapes en sont : choix d'une théorie, formulation d'hypothèses puis vérification sur les données investiguées. Chris Anderson, rédacteur en chef du magazine *Wired*, avait annoncé dans un article provocateur de 2008 la fin de cette méthode, la fin des théories et des modèles scientifiques considérés comme des artefacts subjectifs pour tenter d'expliquer la complexité du réel à notre échelle. Il avait défendu l'avènement de la science des données promouvant la méthode de corrélation des données dans le travail de découverte scientifique : « *The opportunity is great: The new availability of huge amounts of data, along with the statistical tools to crunch these numbers, offers a whole new way of understanding the world. Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all*[26] » (Anderson, 2008).

L'opposition entre corrélation et causalité marque une grande différence épistémologique. La corrélation cherche à donner la primauté à l'action et à la prévision alors que la causalité vise en priorité la compréhension et l'explication des phénomènes étudiés, orientation privilégiée dans de nombreux travaux de sciences sociales. Mais l'approche « traditionnelle » associée à une méthode déductive partant d'une théorie, d'un modèle choisi, laisse peu de place à la découverte de nouveaux cadres de raisonnement. La corrélation et la *data science* privilégient, quant à elles, une méthode inductive conduisant très souvent à formuler des hypothèses avec des boucles rétroactives entre induction et déduction plutôt qu'à partir d'un cadre *a priori.*

Conclusion

L'hybridation du travail scientifique avec l'environnement des données s'est jusqu'à présent située essentiellement au niveau du matériau d'entrée en sciences humaines et sociales. Mais le régime de la donnée massive ouvre aussi de nouvelles pistes avec ses enjeux, ses limites et ses biais. Aucune méthodologie scientifique ne peut de toute façon prétendre à aucun biais. Se saisir de cette nouvelle voie peut être intéressant pour des recherches en SHS. Nous rejoindrons ici la position de Dominique Boullier (2015) qui avait proposé de considérer une 3e génération de SHS sans pour autant se substituer aux autres générations marquées chacune par des types de données, des méthodes et des outils spécifiques. Les méthodes de travail ne sont pas incompatibles, leur complémentarité est certainement fructueuse dès lors que l'on s'écarte d'une posture dogmatique.

26 « L'opportunité est grande : la nouvelle disponibilité d'énormes quantités de données, ainsi que les outils statistiques pour analyser ces quantités, offrent une toute nouvelle façon de comprendre le monde. La corrélation remplace la causalité, et la science peut avancer même sans modèles cohérents, sans théories unifiées ou aucune explication mécaniste. »

Références

Anderson C., 2008, « The end of theory: The data deluge makes the scientific method obsolete », *Wired*, 23 juin. https://www.wired.com/2008/06/pb-theory/

Baek C. et Doleck T., 2021, « Educational Data Mining versus Learning Analytics: A Review of Publications From 2015 to 2019 », *Interactive Learning Environments*, 24 juin.

Beaudouin V., 2016, « Retour aux origines de la statistique textuelle : Benzécri et l'école française d'analyse des données », présentation aux *Journées internationales d'analyse statistique des données textuelles (JADT)*, Nice, 7-10 juin, p. 17-27. https://hal.archives-ouvertes.fr/hal-01376938

Berti-Équille L., 2018 [2006], *Qualité des données*, Saint-Denis, Éd. Techniques de l'ingénieur.

Bonneville A., Tucci I., Vion A. et Giglio L., 2021, *Données de la recherche : pratiques et besoins dans un laboratoire pluridisciplinaire SHS*, rapport de recherche, Laboratoire d'économie et de sociologie du travail (LEST), mai. https://hal.archives-ouvertes.fr/hal-03265603v2

Borgman C. L., 2015, *Big data, little data, no data. Scholarship in the networked world*, Cambridge, MIT Press.

Boullier D., 2015, « Vie et mort des sciences sociales avec le big data », *Socio*, 4, p. 19-37. https://doi.org/10.4000/socio.1259

Cabrera F., 2014, *Les données de la recherche en sciences humaines et sociales : enjeux et pratiques. Enquête exploratoire*, mémoire d'études pour le titre de « Chef de projet en ingénierie documentaire », Institut national des techniques de la documentation (INTD), Cnam. https://memsic.ccsd.cnrs.fr/mem_01128394/document

Chartron G., 2018, « L'Open science au prisme de la Commission européenne », *Éducation et sociétés*, 41, p. 177-193. https://doi.org/10.3917/es.041.0177

Chartron G., 2021, « Régime de médiation des "données" en contexte professionnel », communication orale présentée au *22e Colloque international sur le document numérique* (CIDE 22), Paris, 9-10 déc. https://hal.archives-ouvertes.fr/hal-03500937/

Chevallier P., 2018, « Les données au service de la connaissance des usages en ligne : l'exemple de l'analyse des logs de Gallica », *Les Enjeux de l'information et de la communication*, 19/2, p. 57-67. https://doi.org/10.3917/enic.025.0057

Gonçalves Curty R., 2015, *Beyond « Data Thrifting »: An Investigation of Factors Influencing Research Data Reuse In the Social Sciences*, Dissertation, Syracuse University. https://surface.syr.edu/etd/266

InSHS (Institut des sciences humaines et sociales du CNRS), 2021, *Les sciences humaines et sociales et la protection des données à caractère personnel dans le contexte de la science ouverte. Guide pour la recherche, version 2*, févr. 2021. https://www.ouvrirlascience.fr/les-sciences-humaines-et-sociales-et-la-protection-des-donnees-a-caractere-personnel-dans-le-contexte-de-la-science-ouverte-v2/

Joux A. et Bassoni M., 2018, « Le journalisme saisi par les *Big Data* ? Résistances épistémologiques, ruptures économiques et adaptations professionnelles », *Les Enjeux de l'information et de la communication*, 19/2, p. 125-134. https://doi.org/10.3917/enic.025.0125

Kitchin R., 2014, « Big Data, new epistemologies and paradigm shifts », *Big Data & Society*, 1 (1). https://doi.org/10.1177/2053951714528481

Leonelli S., 2015, « What Counts as Scientific Data? A Relational Framework », *Philosophy of Science*, 82 (5), p. 810-821.

Longhi J., 2020, « Explorer des corpus de tweets : du traitement informatique à l'analyse discursive complexe », *Corpus*, 20. https://doi.org/10.4000/corpus.4567

Ouellet M., 2021, « Logique algorithmique et reproduction sociétale : les médiations sociales saisies par les algorithmes », *Tic & société*, 15 (1), p. 1-7. https://doi.org/10.4000/ticetsociete.5600

Prost H. et Schöpfel J., 2019, « Les entrepôts de données en sciences de l'information et de la communication (SIC). Une étude empirique », *Études de communication*, 52, p. 71-98. https://doi.org/10.4000/edc.8604

Rebouillat V., 2021, « Le partage des données vu par les chercheurs : une approche par la valeur », *Les Enjeux de l'information et de la communication*, 22/1, p. 35-53. https://lesenjeux.univ-grenoble-alpes.fr/2021/varia/03-le-partage-des-donnees-vu-par-les-chercheurs-une-approche-par-la-valeur/

Salem A., 1982, « Analyse factorielle et lexicométrie : synthèse de quelques expériences », *Mots. Les langages du politique*, 4, p. 147-168. https://www.persee.fr/doc/mots_0243-6450_1982_num_4_1_1055

Van Hooland S., Gillet F., Hengchen S. et De Wilde M., 2016, *Introduction aux humanités numériques : méthodes et pratiques*, Louvain-la-Neuve, Éd. De Bœck Supérieur.

Zhang J., Wang W., Xia F., Lin Y.-R. et Tong H., 2020, « Data-Driven Computational Social Science: A Survey », *Big Data Research*, 21.