

Outils de recherche en sciences sociales numériques

Chaire de leadership en enseignement des sciences sociales numériques (CLESSN)

2023-09-13

Table of contents

Avant-propos

Ceci est un exemple de citation Adcock and Collier (2001) .

Introduction

1 Comment les données massives affectent-elles les sciences sociales? Changements actuels et quelques réflexions sur l'avenir

L'apparition des données massives (*big data*) dans le paysage technologique représente un de ces cas de plus en plus communs de phénomène hautement technique dont les effets politiques et sociaux sont remarquables. La discussion publique s'est en effet rapidement emparée du sujet, au point de transformer un moment technologique en phénomène social. Les données massives se trouvent ainsi régulièrement présentées dans l'espace public à la fois comme un moyen puissant de développement et d'innovation technoscientifique, de même que comme une menace à la stabilité de certaines normes sociales telles que la confidentialité des informations privées. Il n'est d'ailleurs pas rare que le discours public s'inquiète du danger que poseraient les données massives à la séparation des sphères publique et privée, pourtant centrale à la conception libérale du rôle de la politique qui structure la majorité des débats sociaux, en amalgamant parfois de manière trop rapide l'objet et l'utilisation qui en est faite. Toutefois, ce même discours public s'emporte aussi rapidement à propos des gains technologiques monumentaux réalisés par l'utilisation des données massives.

Dans le domaine des sciences sociales, les avancées dues à l'utilisation des données massives se font de plus en plus fréquentes et l'impact des données massives dans le domaine de la recherche sociale est en ce sens indéniable. Toutefois, d'un point de vue épistémologique, l'utilisation des données

massives en recherche en sciences sociales dans les dernières années laisse plusieurs questions ouvertes dans son sillage.

Comment l'utilisation des données massives change-t-elle la pratique des sciences sociales? Les données massives causeront-elles un changement de paradigme scientifique? Quels impacts auront-elles sur les traditions scientifiques dominantes telles que le béhavioralisme ou l'individualisme méthodologique en sciences sociales?


Ce chapitre ne prétend pas offrir de réponses définitives à ces questions, mais plutôt des pistes de réflexion par le biais d'une introduction critique de certains points relatifs aux impacts des données massives sur la recherche en sciences sociales. Premièrement, nous présentons une conceptualisation des données massives. Deuxièmement, nous nous penchons sur les impacts des données massives en sciences sociales et soulignons tout particulièrement comment elles affectent les enjeux de la *validité* interne et externe dans le domaine des sciences sociales. Finalement, nous explorons quelques pistes de réflexion sur l'avenir des données massives en sciences sociales en analysant quelques changements *épistémologiques* que ces données pourraient potentiellement entraîner.

1.1 Définition des données massives

Ce qui définit les données massives comme concept est souvent mêlé avec le phénomène social qui l'accompagne. Il est toutefois possible de démêler le tout en distinguant trois approches conceptuelles des données massives qui sont décrites dans la Figure 1.1.

1. Premièrement, les données massives représentent une ***quantité importante de points d'information*** qui varient selon la nature, le type, la source, etc. Ici, la distinction est simplement quantitative. Il s'agit d'une première dimension à la définition des données massives.

1.1 Définition des données massives



1. Définition de base	Quantité importante de données dont la nature, le type, la source, etc. varie
2. Définition technique/technologique	Ensemble de <i>pratiques</i> de collecte, de traitement et d'analyse de ces données
3. Définition sociologique	Innovation technique et technologique, de même que les effets sociaux qui l'accompagne

Figure 1.1: image1

- Deuxièmement, d'une perspective technique et technologique, les données massives constituent un ensemble de ***pratiques*** de collecte, de traitement et d'analyse de ces points d'information. Les données massives représentent donc une technique ou une méthode nouvelle de recherche.
- Enfin, d'une perspective sociologique, les données massives représentent un phénomène incorporant à la fois la dimension propre aux ***développements technologiques, ainsi que les impacts sociétaux de ces développements*** — i.e., les risques à la confidentialité des données, les enjeux relatifs au consentement et à l'autorisation de collecte des informations, les innovations en intelligence artificielle, etc. Cette perspective souligne le caractère essentiellement social des données massives.

Dans les domaines scientifiques et technologiques, la définition courante donnée aux données massives intègre des éléments de ces trois niveaux d'analyse en se référant à la composition et à la fonction des données. Premièrement, la *composition* des données massives est généralement conceptualisée comme comprenant « 4V » : le volume, la variété, la vitesse et

la véracité. Cette conceptualisation jouit d'un large consensus scientifique (Chen, Mao et Liu, 2014; Gandomi et Haider, 2015; Kitchin et McArdle 2016). Par ailleurs, plusieurs chercheurs ont élargi cette définition de la composition des données massives en y incluant, par exemple, la variabilité et la valeur des points de données (CITE). Deuxièmement, la *fonction* des données massives comprend les innovations relatives à l'optimisation, à la prise de décision et à l'approfondissement des connaissances qui résultent de leur utilisation. Ces fonctions touchent des domaines sociaux disparates, incluant le souci d'efficacité et de rendement des secteurs privé et public ainsi que la recherche scientifique pure (Gartner 2012).

1.2 Les données massives et les sciences sociales<!--AFC: La structure du chapitre est mêlante. Certaines sous-sections devraient-elles être des sous-sous-sections? Pourquoi la section actuelle est-elle aussi courte, devrait elle inclure plus de choses?-->

Dans le domaine des sciences sociales, les changements causés par l'utilisation des données massives en recherche sont significatifs. Plusieurs n'hésitent d'ailleurs pas à les qualifier de changements de paradigme dans l'étude des phénomènes sociaux (Anderson 2008; Chandler 2015; Grimmer 2015; Kitchin 2014; Monroe et al. 2015). Dans le cas qui nous intéresse, deux dimensions majeures méritent d'être abordées : (1) une première relative à la validité (interne et externe) des données massives et (2) une seconde, plus large, relative au potentiel changement de posture ou d'orientation épistémologique causé par l'utilisation de ces données en recherche.

1.3 La validité de la mesure en sciences sociales

La validité de la mesure constitue une exigence méthodologique centrale à la recherche en sciences sociales. Les scientifiques cherchent effectivement à s'assurer que ce qui est mesuré — par un sondage, une entrevue, un thermostat ou tout autre outil de mesure — constitue bel et bien ce qui est censé être mesuré. Adcock et Collier définissent plus spécifiquement l'application de la validité de la mesure en sciences sociales par le biais de « scores (including the results of qualitative classification) [that] meaningfully capture the ideas contained in the corresponding concept » (2001: 530).

Toutefois, les problèmes liés à la validité de la mesure sont nombreux et ont une importance considérable. Dans l'étude des phénomènes sociaux et humains, la validité de la mesure prend d'ailleurs une complexité supplémentaire du fait que les données collectées par le biais d'une mesure constituent le *produit de l'observation* d'un phénomène, mais non pas le phénomène en soi. Ainsi, lorsque, dans le contexte d'une recherche, on propose de mesurer l'humeur de l'opinion publique (le phénomène en soi) sur un enjeu politique, on utilise généralement un sondage qui a pour fonction de mesurer le pouls d'un échantillon de la population d'intérêt (ce qui est réellement observé). Cependant, ce que ce sondage mesure ne constitue pas tout à fait l'opinion publique elle-même, mais plutôt un segment populationnel qui se veut représentatif de l'humeur de l'opinion publique. Autrement dit, la mesure et les données collectées ne représentent pas le phénomène — l'opinion publique — en soi.

On a déjà mentionné que la validité de la mesure a de l'importance puisqu'elle garantit que ce qui est mesuré représente réellement ce qu'on croit mesurer. Toutefois, pour être plus spécifique, dans une approche positiviste, la validité de la mesure se traduit généralement par une logique de classification des valeurs attribuées aux différentes manifestations distinctes d'un même phénomène. Par exemple, une mesure de la démocratie comme celle proposée par *Freedom House*, fréquemment

utilisée en science politique, classe les libertés civiles et les droits politiques des États du monde par degré afin de construire un index allant d'un autoritarisme complet à une démocratie parfaite. Les scores représentent, dans ce contexte, une mesure artificielle, mais ordonnée et logique, des idées contenues dans le concept de démocratie telles que libertés civiles et droits politiques. On peut ainsi dire que le souci avec la validité de la mesure traverse les connexions entre (1) le phénomène social étudié (la démocratie), (2) son opérationnalisation (via les libertés civiles et droits politiques) et (3) la méthode de mesure utilisée pour observer et classer d'une certaine façon le phénomène et les données qui en découlent (dans le cas de *Freedom House*, des codeurs indépendants).

1.4 La validité des données massives

En ce qui a trait aux données massives, la question de la validité de la mesure constitue un défi nouveau. Les données massives ont en effet comme avantage d'offrir aux chercheurs soit de nouveaux phénomènes à étudier, soit de nouvelles manifestations et nouvelles formes à des phénomènes déjà étudiés. Les données massives permettent donc d'agrandir la connaissance scientifique.

L'étude de King et al. (2013) représente un cas éclairant de phénomène social que l'utilisation des données massives a rendu possible d'étudier. En se basant sur la collecte de plus de 11 millions de publications sur les réseaux sociaux chinois, King et ses collègues ont pu mesurer la censure exercée par le gouvernement chinois sur les réseaux sociaux ont donc pu observer une manifestation inédite de censure massive qui, sans de telles données, serait probablement demeurée mal comprise d'une perspective scientifique. Le nombre de recherches basées sur l'utilisation des données massives similairement innovantes en sciences sociales est par ailleurs en croissance constante (Beauchamp 2017; Bond et al. 2012; Poirier et al. 2020).

1.4 La validité des données massives

Cependant, il faut aussi souligner que les données massives, en raison de leur complexité, peuvent avoir pour désavantage d’embrouiller l’étude des phénomènes sociaux. Les opportunités scientifiques liées aux données massives s’accompagnent en effet de certaines difficultés méthodologiques. Parmi ces difficultés, trois enjeux sont particulièrement cruciaux : (1) la validité interne, (2) la validité externe et (3) la question d’un changement de posture ou d’orientation épistémologique en sciences sociales causé par les données massives.

1.4.1 Validité interne des données massives

Premièrement, les données massives peuvent représenter un défi à la validité interne des études en sciences sociales en rendant ***pragmatiquement difficile l’établissement de mécanismes causaux clairs***. Ce défi est notamment une conséquence du fait que la plupart des données sont présentement issues d’un processus de génération (*data-generating process*) qui est hors du contrôle des chercheur.e.s. Les données massives proviennent en effet habituellement de sources diverses qui sont externes aux projets de recherche qui les utilisent. Elles ne sont pas donc générées de manière aléatoire sous le contrôle des chercheur.e.s.

Un des problèmes liés à cette situation est qu’il est difficile de garantir une source *exogène* de variation par laquelle les chercheur.e.s éliminent l’effet potentiel des facteurs confondants (*confounders*). La distribution aléatoire d’un traitement et d’un contrôle dans une expérience en laboratoire ou sur le terrain représente le standard le plus élevé permettant de fournir cette source exogène de variation.

Pour le dire autrement, le défi de validité interne avec les données massives constitue un enjeu relatif à la qualité des données. Ce n’est évidemment pas un défi propre ou unique aux données massives. Ce défi s’applique également aux autres types de données. Cependant, dans l’état actuel

des choses, le volume et la variété — deux des 4V — des données massives — textuelles, numériques, vidéos, etc. — peuvent miner la qualité de l'inférence causale entre une cause et une conséquence que permet habituellement un processus contrôlé de génération des données. En somme, la validité interne des données massives est une fonction de la qualité de ces mêmes données.

1.4.2 Validité externe des données massives

Deuxièmement, les données massives représentent un défi plus important pour la validité externe des recherches en sciences sociales (Tufekci 2014; Lazer et Radford 2017; Nagler et Tucker 2015). La préoccupation la plus évidente concerne la *représentativité* des données massives collectées. Comme le soulignent Lazer et Radford (2017), la quantité ne permet pas de corriger pour la non-représentativité des données. Les données massives sont ainsi soumises au même problème de biais de sélection que les autres types de données observationnelles, tels un sondage ou une série d'entrevues, traditionnellement utilisés en sciences sociales.

Le cas célèbre de l'erreur de prédiction du *Literary Digest* lors de la campagne présidentielle américaine de 1936 illustre bien ce problème récurrent. Lors de cette campagne, le *Literary Digest* a prédit à tort la victoire du candidat républicain Alf Landon sur le président démocrate sortant Franklin D. Roosevelt, puisque son échantillon de répondants sur-représentait les électeurs plus aisés, traditionnellement plus républicains, au détriment des électeurs moins aisés, plus généralement proches du Parti démocrate. Cette erreur de surreprésentation dans l'échantillon est due au fait que le *Literary Digest* a effectué un échantillonnage basé sur les listes téléphoniques et le registre des propriétaires de voitures, biaisant par le fait même l'échantillon au détriment des électeurs plus pauvres ne possédant pas de téléphone ou d'automobile, mais qui constituaient un électorat favorable à Roosevelt (Squire 1981). Le biais de sélection du sondage a

1.4 La validité des données massives

ainsi sous-estimé le soutien populaire de Roosevelt de plus de 20 points de pourcentage.

Aujourd’hui, l’utilisation des données massives est soumise aux mêmes risques méthodologiques. L’accumulation massive de données ne permet pas de compenser pour la qualité des données. Les données massives, comme les données plus traditionnelles, sont soumises aux conséquences induites par le processus de génération des données (*data generating process*) comme un échantillonnage.

1.4.3 Données expérimentales

La question du *processus de génération* des données est plus claire quand on considère comment les *données observationnelles* et les *données expérimentales* permettent d’effectuer des *inférences* de manière distincte.

Premièrement, les données massives ne peuvent pas résoudre les enjeux liés aux inférences causales ou explicatives (Grimmer, 2015). En effet, le processus de génération de données expérimentales assure idéalement la validité de l’inférence causale sur l’ensemble de la population visée. Cela prend plus spécifiquement la forme d’un processus de génération des données au sein duquel les chercheur.e.s assurent la distribution aléatoire du traitement entre les deux groupes — traitement et contrôle — garantissant par le fait même une source exogène de variation qui permet d’éliminer l’endogénéité entre la variable indépendante (x) et le résidu (e) et qui assure donc que l’effet observé n’est pas dû à une variable confondante.

1.4.4 Données observationnelles

En ce qui a trait aux données observationnelles, il y a deux points importants. Premièrement, des méthodes d’inférence basées sur des approches par design (*design-based methods*) comme une méthode de régression sur discontinuité ou de variable instrumentale peuvent également garantir

des inférences explicatives et causales valides. Elles nécessitent toutefois plusieurs postulats plus restrictifs dont l’objectif est d’imiter ou de recréer, de la manière la plus fidèle possible, une distribution aléatoire du traitement – ce que la littérature appelle un *as-if random assignment* (Dunning, 2008).

Dans un contexte observationnel, les données massives peuvent donc permettre d’augmenter la précision des estimations causales. Effectivement, comme dans un modèle de régression linéaire, plus l’échantillon est grand, plus l’estimation du coefficient causal ou probabiliste est précise. Par exemple, un échantillon large dans un modèle de régression sur discontinuité permet de restreindre la largeur de bande autour du seuil, garantissant ainsi une distribution presque parfaitement aléatoire des données et une validité plus élevée à l’estimation de l’effet causal.

Deuxièmement, un échantillon de données massives observationnelles issues d’une plateforme comme X — anciennement Twitter — ou Facebook peut fournir une *description* plus fine de certaines dynamiques sociales observées sur les réseaux sociaux. Cependant, c’est la manière dont sont collectées les données de cet échantillon de données massives qui garantit la représentativité de l’échantillon — avec pour objectif un biais de sélection = 0 — et non pas la quantité de données. Généralement, le biais d’un échantillon est une conséquence de la non-représentativité des répondants; dans notre exemple, les utilisateurs des médias sociaux ne sont généralement pas représentatifs de la population entière.

Dans un tel cas, des méthodes de pondération sur des données observationnelles peuvent compenser pour la sur- ou la sous-représentativité de sous-groupes dans un échantillon afin d’assurer la validité de l’inférence entre échantillon et population. Les données massives ont ici une importance puisqu’une pondération fiable nécessite une quantité substantielle d’observations. Une pondération *a posteriori* sera donc plus fiable plus l’échantillon est grand. Les données massives ont ainsi une valeur ajoutée afin d’établir des inférences descriptives plus précises et sophistiquées.

1.5 Pourquoi ce qui se passe actuellement mérite-t-il que l'on s'y attarde?

1.4.5 Validité écologique et observation par sous-groupes

Les données massives peuvent aussi jouer d'autres rôles importants relatifs à la validité externe. Premièrement, les données massives facilitent effectivement la validité externe de certaines études en accroissant la validité écologique (*ecological validity*) des tests expérimentaux, c'est-à-dire le réalisme de la situation expérimentale (Grimmer, 2015: 81). En effet, la variété des sources et des formats de données permet aux chercheurs d'imiter plus concrètement la réalité sur le terrain vécue par les participants aux études.

Deuxièmement, la quantité importante de données rend possible l'observation d'effets précis, spécifiques et inédits par sous-groupes (Grimmer 2015: 81). Alors qu'auparavant, la taille réduite des échantillons ne permettait pas d'effectuer des inférences valides pour des sous-groupes de la population — les écarts-types par sous-groupes étaient trop grands, rendant difficile l'estimation précise d'un paramètre comme la moyenne et impossible celle d'un coefficient —, la taille énorme des échantillons de données massives permet aux chercheurs d'estimer des paramètres qui étaient demeurés extrêmement imprécis jusqu'à aujourd'hui. Notre compréhension des phénomènes sociaux s'en trouve par le fait même approfondie de façon considérable.

1.5 Pourquoi ce qui se passe actuellement mérite-t-il que l'on s'y attarde?

Appréhender l'impact actuel des données massives se révèle d'une importance cruciale pour se préparer à l'avenir. Tout d'abord, cela s'avère propice à une prise de décision éclairée. En scrutant comment ces données ont été rassemblées, traitées et interprétées dans le passé, nous pouvons rehausser la qualité des choix que nous effectuons aujourd'hui dans des domaines aussi diversifiés que la santé, l'économie et l'environnement. De

1 Comment les données massives affectent-elles les sciences sociales? Changements actuels et

	Données observationnelles	Données expérimentales
Processus de génération des données	Non contrôlé par le chercheur	Contrôlé par le chercheur
Type d'inférence causale	Locale (LATE) ou populationnelle (ATE)	Populationnelle (ATE)
Méthodes	Approches par design	Distribution aléatoire du traitement
Exemples	Régression sur discontinuité, variable instrumentale	Expérience de terrain, laboratoire

Figure 1.2: image2

surcroît, l'analyse des données massives met en lumière des tendances et des motifs subtils échappant aux ensembles d'informations plus restreints. Ces découvertes pavent la voie à des concepts innovants et à des avancements technologiques répondant aux mutations des besoins sociétaux. D'autre part, la préoccupation grandissante liée à la préservation de la vie privée et à l'éthique requiert une appréhension approfondie des erreurs passées dans la manipulation de ces données massives. Évitant la réitération de telles erreurs, nous pouvons ériger des cadres réglementaires plus responsables et instaurer des pratiques de traitement respectueuses des droits individuels. Somme toute, la compréhension de l'incidence actuelle des données massives offre une opportunité inestimable pour contrecarrer les égarements passés et façonner un avenir où l'utilisation de ces données s'inscrit dans une démarche éclairée, éthique et propice au bien-être de l'ensemble de la société.

1.6 En guise de conclusion : trois questions ouvertes pour le futur

Comme nous venons de le voir, la quantité et la variété nouvelle des données massives permettent à la fois un approfondissement de l'analyse de certains phénomènes et l'ouverture de nouvelles avenues de recherche. Il faut toutefois souligner que d'une perspective non pas seulement méthodologique/technique, mais plutôt *épistémologique*, les données massives représentent une *complexification* de l'analyse des phénomènes en sciences sociales. Cela soulève au moins trois questions d'importance, dont les réponses ne nous sont pas encore accessibles, pour l'avenir de la recherche en sciences sociales : (1) les données massives entrent-elles (partiellement du moins) en conflit avec l'impératif de parcimonie qui caractérise la science moderne?; (2) ces données sont-elles dans la continuité ou représentent-elles une coupure dans la tradition béhavioraliste en sciences sociales (et en science politique tout particulièrement)?; (3) et finalement, de manière reliée, les données massives proposent-elles ou non une manière de dépasser l'individualisme méthodologique qui caractérise les sciences sociales contemporaines?

2 Le monde du libre

« Vous n'avez pas à suivre une recette avec précision. Vous pouvez laisser de côté certains ingrédients. Ajouter quelques champignons parce que vous en raffolez. Mettre moins de sel car votre médecin vous le conseille — peu importe. De surcroît, logiciels et recettes sont faciles à partager. En donnant une recette à un invité, un cuisinier n'y perd que du temps et le coût du papier sur lequel il l'inscrit. Partager un logiciel nécessite encore moins, habituellement quelques clics de souris et un minimum d'électricité. Dans tous les cas, la personne qui donne l'information y gagne deux choses : davantage d'amitié et la possibilité de récupérer en retour d'autres recettes intéressantes. » - Richard Stallman

Cette analogie illustre bien trois concepts au coeur de la philosophie de Richard Stallman, souvent considéré comme le père fondateur du logiciel libre : liberté, égalité, fraternité. Les utilisateurs de ces logiciels sont libres, égaux, et doivent s'encourager mutuellement à contribuer à la communauté. Ainsi, un logiciel libre est généralement le fruit d'une collaboration entre développeurs qui peuvent provenir des quatre coins du globe. Une réflexion éthique est au coeur du mouvement du logiciel libre, dont les militants font campagne pour la liberté des utilisateurs dès le début des années 1980. La Free Software Foundation (FSF), fondée par Richard Stallman en 1985, définit rapidement le logiciel «libre» [free] comme garant de quatre libertés fondamentales de l'utilisateur: la liberté d'utiliser le logiciel sans restrictions, la liberté de le copier, la liberté de l'étudier, puis la liberté de le modifier pour l'adapter à ses besoins

2 Le monde du libre

puis le redistribuer ¹. Il s'agit ainsi d'un logiciel dont le code source² est disponible, afin de permettre aux internautes de l'utiliser tel quel ou de le modifier à leur guise. Puisque le langage machine est difficilement lisible par l'homme et rend la compréhension du logiciel extrêmement complexe, l'accès au code source devient essentiel afin de permettre à l'utilisateur de savoir ce que le fait programme fait réellement. Seulement de cette façon, l'utilisateur peut *contrôler* le logiciel, plutôt que de se faire contrôler par ce dernier (Stallman, 1986).

2.1 Émergence et ascension

Plusieurs situent les débuts du mouvement du logiciel libre avec la création de la licence publique générale GNU³, en 1983, à partir de laquelle va se développer une multitude de programmes libres. Depuis, la popularité des logiciels libres n'a cessé de croître, alors que des dizaines de millions d'utilisateurs à travers le monde utilisent désormais ces logiciels. Parmi les plus populaires, on retrouve notamment le navigateur Firefox, la suite bureautique OpenOffice et l'emblématique système d'exploitation Linux, qui se développe d'ailleurs à partir de la licence GNU. Les logiciels libres ont différents usages (en passant par la conception Web, la gestion de contenu, les systèmes d'exploitation, la bureautique...). Encore une fois, le logiciel libre est avant tout une philosophie, voire un mouvement de société. C'est une façon de concevoir la communauté du logiciel, où le

¹La redistribution doit évidemment respecter certaines conditions précises, dont l'enfreinte peut mener à des condamnations [<http://www.softwarefreedom.org/resources/2008/shareware.html>].

²Pour rester dans les analogies culinaires, le code source est au logiciel est ce que la recette est à un plat: elle indique les actions à effectuer, une par une, pour arriver à un résultat précis. Encore une fois, cette dernière peut-être adaptée, modifiée, bonifiée.

³expliquer ce qu'est GNU en quelques lignes/le modèle collaboratif de développement logiciel initié par le projet GNU

2.2 Principaux avantages et inconvénients

respect de la liberté de l'utilisateur est un impératif éthique central (**reformuler?**) (Williams et al., 2020:26). Si ce mouvement fut d'abord initié par quelques militants dans les années 1980, c'est aujourd'hui un véritable phénomène sociétal: des milliers d'entreprises, d'organisations à but non lucratif, d'institutions ou encore de particuliers adoptent tour à tour ces logiciels, dont la culture globale et les valeurs (entraide, collaboration, partage) s'arriment avec le virage technologique de plusieurs entreprises à l'ère du numérique (**retravailler, mais l'idée est là**). [blablabla]

Il faut garder en tête que le logiciel libre ne rime pas nécessairement avec gratuité. Bien que plusieurs logiciels libres soient téléchargeables gratuitement (**donner des exemples**), il est aussi possible de (re)distribuer des logiciels libres payants (**reformuler, pas clair**). Par ailleurs, aucun logiciel libre n'est réellement «gratuit» dans la mesure où son déploiement et son utilisation nécessitent généralement différents coûts, dont les degrés sont variables en fonction des compétences et de l'infrastructure dont disposent les utilisateurs (coût d'apprentissage, coûts d'entretien, etc.). Enfin, il est important de garder en tête les logiciels libres possèdent eux aussi une licence - cette dernière est d'ailleurs garante des libertés que confèrent les logiciels libres aux utilisateurs.

2.1.1 Logiciel libre et *open source*

“Les deux expressions décrivent à peu près la même catégorie de logiciel, mais elles représentent des points de vue basés sur des valeurs fondamentalement différentes. L'open source est une méthodologie de développement ; le logiciel libre est un mouvement de société.”

2.2 Principaux avantages et inconvénients

La disponibilité du code source et le mode de développement collaboratif du logiciel libre facilitent également le transfert des connaissances, et ce,

2 Le monde du libre

au-delà des frontières. Où qu'ils soient, les institutions, les entreprises et les particuliers peuvent utiliser ces logiciels et les adapter en fonction de leurs besoins respectifs. Par ailleurs, l'accès libre et égal de tous les internautes à l'ensemble de ces connaissances constitue un enjeu majeur pour la vitalité démocratique des sociétés à l'ère du numérique, caractérisées par une surabondance d'information.

Les logiciels libres, parce qu'ils sont souvent moins coûteux (voire téléchargeables gratuitement) et qu'ils démocratisent l'accès à l'information, contribuent à réduire les disparités en termes d'accessibilité aux nouvelles technologies.

Stallman - Lui-même issu du monde de la recherche scientifique. L'esprit même du logiciel libre est très proche ; contribution à la culture globale de partage, d'entraide, etc. que l'on peut retrouver dans le domaine scientifique

3 Les outils de collecte de données

La révolution numérique engendrée par l'émergence du Big Data représente un important défi pour le monde des sciences sociales (Manovich, 2011; Burrows et Savage, 2014). Elle constitue également une opportunité de recherche enrichissante et innovante permettant une compréhension plus accrue des phénomènes sociaux étudiés par la communauté scientifique (Connelly et al., 2016). Cette meilleure compréhension est permise, entre autres, par l'accès à des données massives concernant les trois acteurs clés de la société démocratique: les citoyens, les médias et les décideurs (Schroeder, 2014; Kramer, 2014). Si l'accès à ces données représente un défi éthique et théorique, tel qu'explicité lors des chapitres précédents, elle représente également un défi technique pour les chercheurs.euses voulant exploiter le potentiel et les opportunités offertes par les données massives (Burrows et Savage, 2014). Le chapitre qui suit vise à offrir un portrait de certains outils de collecte de données pouvant être exploitées par les chercheurs.euses en sciences sociales visant à tirer profit de la révolution numérique. À travers ce chapitre, il sera question d'outils permettant de collecter des données de sondages, des données médiatiques, de même qu'une panoplie de données par le biais d'extracteurs. Ce chapitre offre donc un tour d'horizon de certains outils de collecte de données à la disposition des chercheurs.euses qui souhaitent entamer des recherches en sciences sociales numériques.

3.1 Le Big Data et les différents acteurs de la société :

Le champ d'étude de la science politique repose sur l'étude de trois types d'acteurs distincts ayant un impact sur la condition socio-économique et politique d'une société : les décideurs, les médias et les citoyens. La recherche sur les décideurs comprend entre autres l'analyse des politiques publiques, des partis politiques, de stratégies électorales ou encore l'analyse de discours de politiciens ou d'organisations. L'étude des médias repose largement sur le rôle des médias dans la formation des priorités et des jugements des citoyens quant aux enjeux politiques, de même que sur leur capacité d'influencer l'agenda des politiciens. En ce qui concerne les citoyens, le champ d'étude de l'opinion publique se consacre à l'analyse des comportements et des attitudes politiques des individus. De plus, de nombreuses recherches visant à comprendre le rôle des citoyens dans une société démocratique portent sur l'influence de la société civile de même que sur l'effet des mouvements sociaux.

Chacun de ces champs de recherches se voit confronté à une panoplie de défis théoriques et techniques en lien avec l'émergence des données massives. La révolution technologique permet une étude plus approfondie des phénomènes auxquels sont confrontés les différents acteurs de la société démocratique, en raison de l'importante quantité de données accessible aux chercheurs.euses. Toutefois, la collecte de données permettant de mener à terme de telles études peut s'avérer complexe. Pour chacun des trois acteurs démocratiques énumérés précédemment, les sections suivantes énumèrent et expliquent les capacités techniques d'outils permettant aux chercheurs.euses d'accéder à des données massives. Bien que d'autres outils existent et offrent des résultats satisfaisants, les méthodes suivantes sont particulièrement pertinentes dans une optique d'étude des sciences sociales numériques en raison de leur capacités techniques de même que par la relative simplicité de leur utilisation.

3.2 Plateformes de sondages et collecte de données

Malgré certaines différences méthodologiques, toute recherche doit analyser et interpréter des données fiables et de qualité afin d'émettre des résultats (Nayak & K. A., 2019). Notamment lorsqu'il est question d'étudier les citoyens et l'opinion publique, il est nécessaire d'accumuler suffisamment de données auprès d'un échantillon assez grand afin d'inférer des conclusions sur la population.

Une des méthodes couramment utilisées est le sondage, également nommé panel, enquête, questionnaire, etc. Ils peuvent être manuels ou électroniques, et dans le second cas, peuvent être administrés par ordinateur, par courriel ou via le web (Nayak & K. A., 2019). La différence majeure entre les méthodes manuelles et les méthodes numériques réside dans le fait que les premières impliquent un contact direct entre le chercheur et le répondant, tandis que dans le cas des secondes le contact est indirect (Evans & Mathur, 2018). L'arrivée des données massives et des outils numériques offre une panoplie de nouvelles opportunités de collecte de données pour la communauté scientifique. Lorsqu'exécutée manuellement, la collecte de données et la réalisation de sondages peuvent devenir des tâches lourdement fastidieuses, et de facto, demander énormément de ressources pour mener une recherche à grande échelle. C'est pourquoi les technologies du numérique peuvent faciliter cet aspect de la recherche en fournissant des plateformes de sondages et de collecte de données. De plus, les sondages représentaient en 2016 environ 20% du chiffre d'affaires de l'industrie globale du marketing (Evans & Mathur, 2018). Ces chiffres montrent la pertinence de l'acquisition de compétences nécessaires à la formation de sondages, tant dans le monde académique que professionnel. Le numérique permet donc de créer un questionnaire, de cibler une population et de la contacter, d'entreposer les données des répondants pour ainsi les visualiser, le tout à un coût réduit et plus rapidement que s'il avait été conduit manuellement (Nayak & K. A., 2019). Ainsi, les sondages en ligne ont une portée internationale, permettent le suivi de la ligne du temps,

3 Les outils de collecte de données

offrent des options qui contraignent le répondant à répondre à certaines questions et permettent d'utiliser des arbres de logique avancés que les sondages manuels ne permettent pas.

3.2.1 Les principales plateformes web.

Il existe un large éventail de plateformes de sondages et de collecte de données qui peuvent être utiles dans un contexte académique. Cet ouvrage se limite à cinq d'entre elles : Qualtrics, REDCap, SurveyMonkey, Google Forms et Typeform. Cependant, il n'est pas déconseillé de se renseigner sur les autres plateformes disponibles en fonction de ses besoins et de ses ressources. Voici une liste non-exhaustive d'autres plateformes de collecte de données en ligne : LimeSurvey, Zoho Survey, Qualaroo, Formstack, Wufoo, Checkbox Survey, SmartSurvey, QuickTapSurvey, So-GoSurvey, Snap Surveys, AskNicely, Opinio, Alchemer, Cognito Forms, Feedbackify.

3.2.1.1 Qualtrics (<https://www.qualtrics.com/>)

Cette plateforme est une des plus reconnues et utilisées à l'international, tant dans le milieu académique que dans le secteur privé. En plus d'offrir des outils de collecte de données et de sondages, Qualtrics est utilisé dans le marketing et dans la gestion de l'expérience client. Il est donc pertinent de se familiariser avec cet outil, car il offre des compétences pratiques pour la recherche, mais également pour obtenir des opportunités de carrière. Qualtrics offre plusieurs services pratiques pour la collecte de données, avec des options flexibles pour la programmation et l'administration des sondages. Par exemple, Qualtrics s'adapte à différents formats en fonction de l'appareil du répondant (Evans & Mathur, 2018).

3.2 Plateformes de sondages et collecte de données

3.2.1.2 REDCap (<https://www.project-redcap.org/>)

Research Electronic Data Capture (REDCap) permet de construire et de gérer des sondages ainsi que des bases de données. Pour accéder aux services de cette application, il est nécessaire d'être un partenaire du REDCap Consortium ou membre d'une organisation qui en fait partie. Seules les organisations à but non lucratif peuvent adhérer au Consortium. Les données et les sondages qui y sont produits peuvent être partagés et utilisés par différents chercheurs issus de diverses institutions. L'exportation vers différents types de fichiers (Excel, PDF, SPSS, SAS, Stata, R) est possible. Ce qui distingue REDCap des autres applications est sa compatibilité avec les dossiers médicaux, sa sécurité pour les données sensibles, ainsi que son approche académique à la collecte de données par sondage.

3.2.1.3 SurveyMonkey (<https://www.surveymonkey.com/>)

SurveyMonkey se distingue des autres applications en permettant de construire et gérer des sondages/formulaires à l'aide d'une interface conviviale sans toutefois perdre de ses fonctionnalités. En plus d'avoir recours aux nouvelles technologies de l'I.A. pour aider à construire des sondages adaptés à vos besoins, cette application propose plusieurs centaines de modèles personnalisables élaborés par des experts dans le domaine. SurveyMonkey permet également l'analyse des données et la création de rapports directement sur l'application, en plus de permettre l'exportation vers d'autres types de programmes. Les forfaits varient en gamme de tarifs, allant du gratuit avec des fonctionnalités restreintes, jusqu'aux options payantes destinées aux particuliers et aux entreprises.

3.2.1.4 Google Forms (<https://docs.google.com/>)

Cette application se distingue par sa simplicité et son accessibilité, en grande partie grâce à l'omniprésence de google tant dans le monde académique que dans la vie courante. Google Forms est inclus dans le forfait de base du Google Workspace, ce qui le rend largement compatible avec les autres applications de Google, en plus d'être disponible gratuitement. Bien que ses fonctionnalités soient moins avancées que celles de ses concurrents, Google Forms peut convenir pour des sondages plus simples et rapides grâce à son interface conviviale, à sa fonction d'analyse de données directement sur la plateforme, ainsi qu'à ses modèles préfabriqués.

3.2.1.5 TypeForm (<https://www.typeform.com/>)

Si votre objectif est de produire des formulaires avec une esthétique attrayante, moderne et interactive, TypeForm est la plateforme idéale. Elle permet de se concentrer sur l'expérience de l'utilisateur et de l'impliquer dans le sondage grâce à son aspect visuel. Cette plateforme dispose d'une option gratuite, ainsi que plusieurs forfaits payants. Typeform est également compatible avec plusieurs applications de gestion du flux de travail (Zapier, Google Sheets, Slack, etc).

3.2.2 Les limites des sondages en ligne

Néanmoins, les sondages en ligne comportent des défis, notamment en ce qui concerne l'échantillonnage, les taux de réponse et les caractéristiques des non-répondants. Il est également nécessaire de se méfier des enjeux

3.2 Plateformes de sondages et collecte de données

éthiques et de confidentialité (Nayak & K. A., 2019). Comme la généralisation essentielle pour conférer une valeur scientifique à ses résultats de recherche, les sondages en ligne ont leurs limites. En effet, il est crucial de connaître la population cible pour effectuer des inférences fiables, et l'échantillonnage doit reposer sur des caractéristiques précises. Même si des informations démographiques peuvent être collectées et des quotas utilisés, il n'est toutefois pas réellement possible de confirmer les informations sur le répondant (Andrade, 2020). Les sondages traditionnels où l'on retrouve un contact direct sont plus susceptibles de permettre de brosser un portrait plus complet du répondant (Evans & Mathur, 2018). Les répondants avec des biais peuvent également plus facilement répondre aux sondages en ligne et limiter la généralisation (Andrade, 2020). Les sondages en ligne sont également souvent perçus comme des pourriels, ont généralement de faibles taux de réponse, sont impersonnels et peuvent avoir des instructions peu claires. Ils ont également leurs lots d'enjeux de confidentialité (Evans & Mathur, 2018).

Conseils méthodologiques à la réalisation d'un sondage numérique
(Evans & Mathur, 2018)

L'article de Evans et Mathur (2018) est une revue de littérature observant l'évolution des sondages numériques depuis la parution de leur dernier article sur le sujet en 2005. À travers cet article, les auteurs offrent des conseils méthodologiques en fonction de leur analyse de contenu de la littérature scientifique. Les conseils d'Evans et Mathur (2018) sont résumés ci-dessous. Afin d'obtenir plus de détails, n'hésitez pas à approfondir votre lecture de l'article. De plus, bien qu'il s'agisse d'un article crédible et largement documenté, il est toujours pertinent de consulter des sources spécifiques à vos besoins.

1. Définir le but du sondage avant la méthodologie. Lorsque possible, inclure des hypothèses testables et des méthodes basées sur des fondations théoriques.
2. Choisir le type de sondage.

3 Les outils de collecte de données

3. Décider des méthodes d'échantillonnage, des quotas et des échéances.
4. Déterminer le responsable de la construction du sondage.
5. Soyez transparent en divulguant le but du sondage, la façon dont les données seront utilisées ainsi que l'auteur du sondage.
6. Les questions et les catégories de réponses doivent être élaborées de manière objective et dans une perspective de convivialité.
7. Les sondages doivent être assez légers pour favoriser un taux de réponse positif, mais assez complet pour avoir l'information nécessaire.
8. Ils doivent également être attrayant afin de favoriser leur complétion par le répondant.
9. S'assurer de l'anonymat du répondant.
10. Il faut régulièrement procéder à des tests afin de corriger les faiblesses du questionnaire.
11. Déterminer qui administre le sondage, qui collecte l'information, et qui analyse les données.
12. Établir un échéancier pour les différentes étapes de l'étude.
13. Suite à la collecte de données, entreposer les données brutes dans un fichier électronique.
14. Utiliser les méthodes appropriées (qualitatif ou quantitatif), et analyser les données selon les buts de l'étude.
15. Dans le cas d'une recherche académique, il est important d'avoir une section dédiée aux limites de l'étude.
16. Conserver l'anonymat des répondants lors de l'analyse et de la publication.
17. Agir sur les résultats. Rien ne sert de conduire un sondage qui ne contribue pas à la croissance du savoir ou n'apporte pas de changement stratégique ou organisationnel.
18. Toujours se plier à un code d'éthique rigide.

Il s'agit donc ici d'un court résumé des plateformes de sondages et de la collecte de données en ligne, tentant de couvrir l'essentiel de cet outil afin de vous aider lors de votre parcours académique, ou simplement comme aide-mémoire pour la réalisation d'un sondage. Les outils énumérés précédemment permettent une étude approfondie de phénomènes concernant les

3.3 Factiva : outils de récolte de données médiatiques

citoyens. Bien sûr, il n'est pas possible de couvrir l'entièreté de cet outil très complexe et ayant évolué dans le temps, cette section ne sert donc que de point de départ si vous vous intéressez à l'élaboration d'un sondage numérique, ou si vous avez besoin de récolter des données. Il vous est donc recommandé de vous renseigner davantage avec d'autres ressources afin de compléter ce qui est indiqué dans cet ouvrage.

3.3 Factiva : outils de récolte de données médiatiques

L'émergence de nouvelles technologies de même que la fragmentation médiatique, causée notamment par l'apparition de chaînes de nouvelles en continu, ébranlent considérablement les écosystèmes médiatiques occidentaux (Chadwick, 2017). Un récent courant de recherche se penche sur le rôle des médias relativement aux comportements des individus dans une perspective de fragmentation médiatique. Ces changements de dynamique médiatiques permettent aux individus de choisir leurs sources d'information. Cette fragmentation aurait conséquemment pour effet de contribuer à la formation de chambres d'écho. Ainsi, les études sur les effets des médias visent à comparer les agendas de différentes organisations médiatiques de même que de comprendre le cadrage de la nouvelle qu'ils offrent aux citoyens. Pour effectuer de telles études comparées, l'accès à des données médiatiques est essentiel. L'arrivée de données massives permet de nouvelles avenues de recherche pour les chercheurs.euses en sciences sociales en raison de l'importante quantité de données accessibles aux chercheurs.euses, ce qui permet une compréhension accrue des réalités médiatiques modernes.

L'outil Factiva offre un accès à l'ensemble des articles d'une panoplie de médias provenant d'une vaste sélection de pays. Le moteur de recherche est opéré par Dow Jones et offre également l'accès à des documents d'entreprises. En revanche, l'accès qu'il offre aux contenus médiatiques

3 Les outils de collecte de données

est particulièrement pertinent pour la communauté scientifique en communication et en sciences sociales. Il offre l'accès à plus de 15 000 sources médiatiques provenant de 120 pays. Il permet de télécharger une quantité illimitée de documents RTF, un format de fichier de texte, pouvant contenir jusqu'à 100 articles chacun. En outre, ils peuvent être sélectionnés automatiquement en cochant le bouton proposant de sélectionner les 100 articles de la page de résultat. Chaque page de résultat contient 100 articles à la fois. Enfin, Factiva permet également de filtrer les doublons.

En outre, cet outil permet également de lancer une requête de recherche par mots-clés et par date qui permet, par exemple, de récolter les articles médiatiques concernant un sujet précis dans une ligne de temps déterminée. De manière plus précise, Factiva permet de filtrer la recherche d'articles par source, par date, par auteur, par sociétés, par sujet, par secteur économique, par région et par langue. Disons qu'un.e chercheur.euse désire comparer la couverture médiatique d'une élection donnée. Il peut, par le biais de Factiva, sélectionner tous les articles contenant le mot « élection » dans une sélection de médias, et ce, durant la période de l'élection. Les mots clés sélectionnés peuvent être adaptés aux désirs de la personne chercheuse de manière à inclure des mots qui peuvent être mis ensemble ou à un maximum d'intervalle de mot. L'utilisation des signes « and » et « or », aussi connus sous le nom d'opérateurs booléens, permettent d'ajouter un mot dans la requête de recherche. En ajoutant near5, l'on peut spécifier qu'il doit y avoir un maximum de 5 mots entre les deux mots recherchés. L'on peut également mettre certains signes à la fin de mots, ce qui permet de préciser le champ de recherche. Par exemple, dans une étude récoltant des articles sur les immigrants, le mot immigrant pourrait être écrit de la manière suivante : `immigra*`. Ainsi, tous les mots débutant par ce suffixe seraient inclus de la recherche d'article, ce qui comprend donc : `immigrant`, `immigration`, `immigrants`, `immigrante`, etc. La Figure 1 est une capture d'écran de l'interface de recherche de Factiva. Ainsi, en ajoutant un opérateur booléen, l'on peut préciser un champ de recherche.