

Outils de recherche en sciences sociales numériques

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

2023-10-08

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. varient |
| 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_1

- 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.
- Finally, 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 surreprésentait 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 sur-

| | 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__2

croît, 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

”Vers une science numérique plus transparente: l’apport du logiciel libre et du code ouvert dans les sciences sociales” author: ”Catherine Ouellet et Jozef Rivest”

Catherine Ouellet et Jozef Rivest

Ce chapitre vise à initier les lecteurs et lectrices aux concepts fondamentaux du logiciel libre. Pour ce faire, nous présenterons, dans un premier temps, l’historique de ce mouvement afin de pouvoir le situer temporellement. De cette façon, nous pourrions mieux comprendre les motivations derrière ce mouvement, mais aussi ses influences actuelles. Ensuite, nous distinguerons le logiciel libre du code ouvert. Bien que les deux soient très près l’un de l’autre, il est important de les distinguer puisqu’ils ne renvoient pas aux mêmes caractéristiques et aux mêmes fondements. Après coup, nous utiliserons un exemple concret pour illustrer le propos: R, et ses différentes librairies. La dernière section du chapitre présentera certains avantages, certains inconvénients ainsi que des défis qui se posent pour ce monde. En guise de conclusion, nous souhaitons mettre l’emphase sur l’apport du logiciel libre et du code ouvert afin d’assurer la transparence, la reproductibilité ainsi que la qualité des recherches scientifiques.

« 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

2 Le monde du libre

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 (Williams, Stallman, and Masutti 2010)

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 étant 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 puis le redistribuer¹[La redistribution doit évidemment respecter certaines conditions précises, dont l'enfreint peut mener à des condamnations [http://www.softwarefreedom.org/resources/2008/shareware.html]. Il s'agit ainsi d'un logiciel dont le code source¹ 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

¹Pour 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.

(Stallman 1986)

2.1 Émergence et sémantique du *libre*

Plusieurs situent les débuts du mouvement du logiciel libre avec la création de la licence publique générale GNU, en 1983, à partir de laquelle va se développer une multitude de programmes libres. 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. Aujourd’hui, il s’agit d’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. Les logiciels libres ont différents usages, en passant par la conception Web, la gestion de contenu, les systèmes d’exploitation, la bureautique, entre autres. Ils permettent donc de répondre à plusieurs types de besoins numériques et informatiques.

“Les principes du logiciel libre ont également inspiré de nombreuses initiatives non directement liées à l’informatique et au développement des logiciels libres. La plus connue est sans aucun doute Wikipédia, qui se définit comme une encyclopédie libre, s’inspirant en cela explicitement du modèle du logiciel libre. Soulignons également les licences Creative Commons et le mouvement des archives ouvertes et de libre accès aux revues scientifiques”

Attention, 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 respect de la liberté de l’utilisateur est un impératif éthique (Williams, Stallman, and Masutti 2010). Par conséquent, le terme libre, *free* en anglais, porte à confusion. Celui-ci ne signifie pas qu’un logiciel libre est

nécessairement gratuit. Certes, plusieurs sont effectivement téléchargeables gratuitement. Toutefois, il est aussi possible de (re)distribuer des logiciels libres payant. 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 que les logiciels libres possèdent eux aussi une licence - cette dernière est d'ailleurs garante des libertés que confèrent les logiciels libres aux utilisateurs.

2.2 Logiciel libre et code ouvert

Parallèlement au logiciel libre, il y a aussi le code ouvert, ou *open source*. *A priori*, la dénomination du logiciel libre et celle de l'*code ouvert* semble suggérer qu'il s'agit de synonymes. Dans les deux cas on dirait que l'on fait référence à des logiciels, par exemple, qui sont exempts de restrictions d'utilisations et auxquels les utilisateurs peuvent participer au développement. Cependant, il y a une distinction importante entre les deux.

Bien que les deux renvoient sensiblement aux mêmes types de logiciels, les tenants de ces approches ne partagent pas la même perspective. Comme Stallman (2022) l'explique, le logiciel libre est d'abord et avant tout un mouvement qui fait “campagne pour la liberté des utilisateurs de l'informatique”. Le code ouvert, quant à lui, met l'accent sur les avantages pratiques, plutôt que de militer pour des principes.

Le terme *code ouvert* sera introduit seulement en 1998 afin de clarifier l'ambiguïté dans la dénomination “logiciel libre”², *free software* en anglais, afin de spécifier que le code source était accessible, et non pas que le logiciel était “gratuit” (Ballhausen 2019). De plus, les logiciels code ouvert,

²Soit ceux qui ont été conçus suivant les principes philosophiques et “moraux” qui sous-tendent ce mouvement.

doivent respecter certains critères quant à la distribution de leurs logiciels (Open Source Initiative 2006). Nous aborderons ces critères dans le prochain paragraphe.

Afin de mieux distinguer les deux, il est utile de faire référence aux critères qui composent ces deux éléments, et qui constituent la base de leur définition. Tout d’abord, le logiciel libre se définit sur la base de quatre libertés: 1) liberté d’utiliser le programme tel que désiré; 2) liberté d’étudier le fonctionnement du programme et de le modifier pour ses propres besoins; 3) liberté de re-distribuer des copies; 4) liberté de distribuer des copies de la version “améliorer” du programme pour ses pairs (Ballhausen 2019). Concernant le *code ouvert*, tout logiciel qui souhaite être inclut sous cette appellation doit respecter dix critères: 1) Redistribution gratuite; 2) doit inclure le code source; 3) doit permettre les modifications et les travaux dérivés; 4) intégrité du code source; 5) ne doit pas discriminer des personnes et/ou groupes; 6) ne doit pas restreindre personne dans l’utilisation du logiciel pour un domaine d’activité; 7) distribution d’une licence pour l’utilisation; 8) la licence ne doit pas être spécifique pour un produit; 9) la licence ne doit pas placer de restriction sur d’autres programmes; 10) la licence doit être technologiquement neutre³ (Open Source Initiative 2006).

Il est aussi utile de les distinguer des logiciels “non-libres”, soit les logiciels propriétaires: “Son utilisation, sa redistribution ou sa modification sont interdites, ou exigent une autorisation spécifique, ou sont tellement restreintes qu’en pratique vous ne pouvez pas le faire librement” (Système d’exploitation GNU 2023). Par contraste, la licence libre confère des droits de propriétaire. L’utilisateur a le droit d’installer le logiciel sur autant d’ordinateurs que désiré, le modifier selon ses besoins et le distribuer avec ou sans ses modifications. Il peut même demander d’être payé pour distribuer des copies, avec ou sans ses modifications. Par exemple, le logiciel Ubuntu, une version de Linux, peut être téléchargé gratuitement du

³Pour plus d’informations sur ces caractéristiques, nous encourageons les lecteurs à se référer au lien web de la source. Ils y trouveront un contenu détaillé pour chacune des caractéristiques sus-mentionnées.

2 Le monde du libre

site Ubuntu.com. Il est aussi vendu par Amazon.com pour 12\$ la copie, plus les frais d'expédition!

Comme nous le constatons, le logiciel libre et le *code ouvert* ont certaines similitudes puisqu'ils adhèrent tous les deux à la même vision du logiciel, ainsi que de son accessibilité. Toutefois, il est important tout de même de les distinguer puisqu'ils ont des origines différentes, et qu'ils mènent à certaines pratiques qui sont différentes. La prochaine section utilise un cas concret afin d'expliquer l'effet du libre, et l'utilité que cela peut avoir.

2.3 Cas d'étude: R

Afin d'illustrer le tout plus concrètement, nous utiliserons ici le cas du logiciel R. Il s'agit d'un logiciel statistique que tous les utilisateurs peuvent télécharger gratuitement, et dans lequel il n'y a pas d'achats supplémentaires pour avoir accès à des fonctionnalités supplémentaires par exemple. Bien que ce logiciel soit déjà riche en fonctions et commandes, plusieurs utilisateurs ont développé des *packages*, des librairies externes, afin de bonifier les fonctions de base (Arel-Bundock 2021). Utilisons un cas d'étude afin de démontrer l'apport des librairies externes. Par exemple, je souhaite savoir la probabilité de survie à bord du Titanic en fonction du genre. Je pourrais résumer mon intérêt avec sous la notation suivante: $P(Y = \text{Survie} | X = \text{Femme})$. Cela se lit "la probabilité de survie étant donné que nous soyons une femme". Pour ce faire, je dois utiliser l'ensemble de données `titanic`, disponible en format csv. Je dois donc installer et télécharger la librairie `readr` afin que R puisse importer et lire les données. Ensuite je vais utiliser la commande `table`, offerte dans celles de base, afin de visionner mes données. Cette dernière commande affichera un tableau croisé.

```
library(readr)
```

①

```
dat <- read_csv("data/titanic.csv") ②
```

```
table(dat$survie, dat$femme) ③
```

- ① Téléchargement de la librairie **readr** qui nous permettra de lire des ensembles de données en format **.csv**.
- ② Importation d'une banque de données en format **.csv**.
- ③ Impression d'un tableau croisé afin d'observer la distribution des hommes et des femmes (colonnes), croisé avec la survie (lignes).

```

      0    1
0 709 154
1 142 308

```

Comme nous le voyons ici, la librairie **readr**, développé par plusieurs individus⁴, nous a permis d'importer l'ensemble de données sur le Titanic. Toutefois, le format du tableau n'est pas très esthétique. Pour remédier à ce problème, nous pouvons installer et utiliser la librairie **modelsummary** qui nous permettra de créer rapidement des tableaux croisés plus esthétique, et qui contiendront davantage d'informations, facilitant la lecture et notre compréhension de la relation qui nous intéresse.

```
library(modelsummary) ①
```

```
Tableau.2 <- datasummary_crosstab(survie ~ femme, data = dat) ②
```

```
Tableau.2
```

- ① Téléchargement de la librairie **modelsummary**.

⁴Pour avoir la liste complète des contributeurs, les lecteurs peuvent utiliser la commande **?readr** dans **R**, ou bien consulter le lien suivant <https://readr.tidyverse.org>

2 Le monde du libre

② Création d'un tableau croisé à l'aide de la commande `datasummary_crosstab()`.

| | survie | 0 | 1 | All |
|-----|--------|------|------|-------|
| 0 | N | 709 | 154 | 863 |
| | % row | 82.2 | 17.8 | 100.0 |
| 1 | N | 142 | 308 | 450 |
| | % row | 31.6 | 68.4 | 100.0 |
| All | N | 851 | 462 | 1313 |
| | % row | 64.8 | 35.2 | 100.0 |

Comme nous le voyons, la commande `datasummary_crosstab()` permet facilement de créer des tableaux non seulement plus esthétiques, mais aussi plus informatif. C'est très utile si l'on souhaite incorporer des tableaux dans notre rapport finale, surtout que cette commande nous permet d'exporter les tableaux sous différents format (.docx, **LaTeX**, .qmd, etc.)

Ces deux librairies que nous venons de présenter en exemple, ne sont que deux des 19 897 disponibles pour **R**. Elles illustrent très bien la contribution que les utilisateurs peuvent faire au logiciel. Surtout, ces *add on* ont été développés de manière bénévole. Les contributeurs le font par “passion”, et pour en faire profiter la collectivité d'utilisateurs.

Les logiciels libres permettent aux utilisateurs de jouir d'une plus grande liberté dans leur utilisation, ce qui génère des externalités positives puisque ces gens peuvent créer de nouvelles commandes ou fonction et en faire bénéficier toute la collectivité. L'exemple que nous avons utilisé avec **R** ici reflète très bien cet avantage. La prochaine section de ce chapitre se penche plus en profondeur sur les autres avantages ainsi que sur les inconvénients de ces logiciels.

2.4 Principaux avantages et inconvénients

Dans cette section, nous ne dresserons pas un portrait exhaustif de tous les avantages et les inconvénients du logiciel libre. Cette tâche serait fastidieuse et peu intéressante pour les lecteurs. Notre but ici est de présenter les certains avantages qui sont propre aux logiciels, ainsi que les inconvénients de ceux-ci.

2.4.1 Avantages

2.4.1.1 Le partage et co-construction des connaissances

La grande liberté que ce type de logiciel offre favorise la collaboration entre les utilisateurs, et ce à une échelle pouvant être internationale. Les interactions entre les chercheurs créent une dynamique d'« innovation ascendante » et d'entraide (Couture 2014). Ce résultat constitue un important avantage pour le développement de ces logiciels. Selon certains, et comparativement aux logiciels privés, les logiciels libres ont un niveau plus élevé d'innovation (Smith 2002). Contrairement à ceux qui se développent de manière privée et fermée, les logiciels libres permettent à tous les utilisateurs de participer au développement. Ceux-ci partagent ensuite leurs améliorations, ce qui stimulent à son tour de nouvelles initiatives. Ainsi, un certain savoir est généré dans cette situation. De plus, il est raisonnable de penser que l'utilité des améliorations, ainsi que leur utilisation par les utilisateurs en fonction de leur besoin, comme dans le cas de la recherche sociale avec R permet de générer un savoir collaboratif (Couture 2020). Amélioration constante, entraide, savoir partagé et plusieurs milliers de contributeurs (Couture 2014), ces éléments résument très bien la philosophie du logiciel libre.

Comme nous le verrons dans la section suivante, cet avantage est couplé avec ceux économiques. Les bas coûts démocratisent l'accès à plusieurs logiciels qui sont utiles pour mener des analyses scientifiques. Et ce, pour tous les utilisateurs dans le monde.

2.4.1.2 Avantages économiques : Une plus grande accessibilité pour tous

Nous pouvons aussi mentionner certains avantages économique dans l'utilisation de logiciels libres. Le principal avantage économique des logiciels libres est son faible d'acquisition et de renouvellement pour les particuliers. Cet avantage individuel génère plusieurs externalités positives.

Tout d'abord, certains logiciels statistiques et programmes informatiques, tel que Stata et SPSS, coûtent plusieurs centaines, voir des milliers de dollar. De plus, la licence doit être renouvelée annuellement, ce qui limite l'accès à ces logiciels. Comparativement, pour les logiciels libres, la licence d'acquisition coûte bien souvent moins cher, et aucun renouvellement de licence n'est demandé dans la plus part des cas. Étant donné que les chercheurs doivent souvent faire face à des contraintes budgétaires, les logiciels libres deviennent des outils intéressant afin de minimiser les coûts de la recherche (Yu and Muñoz-Justicia 2022). Avantage encore plus important pour les chercheurs dans les pays du Sud global (Santillán-Anguiano and González-Machado 2023). L'accessibilité de ces ressources permet donc de réduire l'écart dans la production scientifique entre les pays du Sud et ceux du Nord. De plus, elle permet à tous de bénéficier d'outils pédagogiques accessibles, ce qui favorise l'acquisition ainsi que le développement de compétences méthodologiques.

Dans le cadre d'une formation universitaire, il peut être pertinent d'enseigner aux étudiants à se servir de logiciel statistique ou d'analyse de texte. L'acquisition de ces compétences peut être précieux tant pour ceux et celles qui souhaitent se diriger vers le milieu académique, que pour ceux et celles qui visent le marché professionnel. D'ailleurs sur le site web de la banque d'emplois du gouvernement du Canada⁵, les conditions d'emplois

⁵Ces informations proviennent du site web suivant:
<https://www.jobbank.gc.ca/marketreport/outlook-occupation/17882/ca>

2.4 Principaux avantages et inconvénients

sont en ce moment⁶ très bonnes, et une pénurie de main d'oeuvre est anticipé dans ce secteur entre 2022-2031. Ces compétences sont d'autant plus précieuses aujourd'hui, dans le monde de données dans lequel nous vivons.

Ensuite, le logiciel libre est adaptable et modifiable. Ces coût techniques de développement restent néanmoins nettement inférieurs aux coûts de renouvellement et de mise à jour des logiciels propriétaires dans bien des cas. L'argent sauvée des licences peut alors être investie dans le développement du logiciel libre (Béraud 2007). Cependant, une transition vers les logiciels libres ne doit pas se faire seulement sur des bases économiques, mais dans une perspective globale de changement de cultures. Changer pour des raisons purements économiques viendrait à violer l'essence même de la philosophie du logiciel libre, qui se veut davantage être un esprit de collaboration et de transparence. Par conséquent, il est important d'incorporer aussi les valeurs et la philosophie dans notre utilisation

Pour résumer, les logiciels libres permettent donc une plus grande égalité dans l'accès aux nouvelles technologies, puisqu'ils ont dans la majorité des cas, des coûts d'acquisition nettement moindre. (Oui et non, l'acquisition financière est une chose, mais il y a d'autres barrières à l'utilisation tel que l'apprentissage à faire pour apprendre un langage de programmation, l'achat de matériel informatique, etc.) Cependant, considérant cela, donner l'exemple de l'étude qui montre que c'est beaucoup plus économique, même si l'on doit compter les coûts de formation, le soutien technique, l'entretien et la maintenance. Karjalainen (2010).

⁶En date d'écire ces lignes, septembre 2023.

2.4.2 Inconvénients et défis:

2.4.2.1 Coûteux en temps

Dans leur texte, Paura and Arhipova (2012) soulèvent une critique faite envers certains logiciels libres, notamment envers R. Le problème principal d'enseigner les statistiques avec des logiciels libres est qu'ils sont compliqués à apprendre ainsi qu'à utiliser; par conséquent, les étudiants passeraient plus de temps à tenter de résoudre les erreurs de programmation plutôt que d'apprendre les statistiques. Il est vrai que ces logiciels demandent un investissement en temps, afin d'être en mesure de mener ses propres analyses statistiques. Par exemple, R demande l'apprentissage d'un langage de programmation afin de pouvoir utiliser le logiciel à son plein potentiel.

La syntaxe de certaines libraries demandent aussi un certain temps d'adaptation. Par exemple, je souhaite recoder la variable `femme`, de l'ensemble de données `titanic`, afin de remplacer les valeurs numériques actuelles (0, 1) par des valeurs nominales (homme, femme). La section de code ci-dessous réalise cette tâche avec les commandes de base de R et celle du `tidyverse`.

```
dat$femme[dat$femme == 0] <- "Homme" ①  
dat$femme[dat$femme == 1] <- "Femme"  
table(dat$femme) ②
```

- ① Utilisation de la commande de base dans R pour recoder la variable `femme`.
- ② Vérification que la manipulation a bien fonctionné.

```
Femme Homme  
462    851
```

2.4 Principaux avantages et inconvénients

```
library(tidyverse) ③
dat$femme <- recode(dat$femme, `0` = "Homme", `1` = "Femme") ④
table(dat$femme) ⑤
```

- ③ Téléchargement de la librairie `tidyverse`.
- ④ Recodage de la variable `femme` à l'aide de la commande `recode`.
- ⑤ Vérification que la manipulation a bien fonctionné.

```
Femme Homme
462      851
```

Toutefois, l'orsque l'on compare le coût d'apprentissage avec les bénéfices tirés, il est plus difficile de soutenir qu'il s'agit d'un désavantage. L'habileté que nous développons devient très utile par la suite, puisqu'elle nous permet de manipuler ainsi que d'analyser des données. Surtout, ces compétences s'inscrivent dans la longue durée, alors que l'apprentissage est plutôt de courte à moyenne durée. Surtout, la logique derrière la syntaxe de base de R et celle d'une nouvelle librairie reste sensiblement inchangée. Par conséquent, lorsque nous avons une bonne compréhension du fonctionnement de base de R, l'apprentissage d'une nouvelle librairie se fait relativement rapidement. Certaines, comme `dplyr` du `tidyverse` facilite grandement la manipulation des données comparativement aux commandes de base.

Pour résumer, bien que l'apprentissage d'un langage de programmation demande un investissement en temps, les bénéfices générées par ces nouvelles compétences dépassent le coût initial.

2.4.2.2 Problème de transparence

L'arrivée des sciences informatiques a fait émerger des problèmes de reproductibilité des protocoles scientifiques (Janssen 2017). Le problème prin-

2 Le monde du libre

cial est relatif à l'accès au code utilisé par les chercheurs. Par exemple, il est possible de réaliser des analyses statistiques avec R sans partager le code utilisé, ce qui limite la transparence du processus scientifique. Dans cette situation, il est difficile de savoir si des erreurs de codage ont été commises, volontairement ou involontairement, affectant ainsi les résultats partagés.

Afin de remédier à ce problème, certains logiciels tel que GitHub⁷ participent à la transparence des résultats scientifiques (Fortunato and Galassi 2021). Ce logiciel permet aux chercheurs de partager leur code afin qu'il puisse être accessible pour tous. Il est important de mentionner ici que l'installation et la configuration de GitHub peut s'avérer difficile pour ceux et celles qui ne sont pas initiés à l'informatique. Cela constitue une certaine barrière dans l'utilisation de ce logiciel. Toutefois, nous souhaitons tout de même présenter l'utilité de ce logiciel puisqu'il permet de rendre les processus ainsi que les résultats de recherche plus transparents.

Par exemple, si l'on réalise une analyse statistique de la relation entre l'économie et le vote, nous pourrions partager l'ensemble du code que nous avons utilisé sur GitHub. D'une part cela permettrait aux utilisateurs de vérifier si les résultats sont honnêtes, et d'autre part de réutiliser le code pour mener leurs propres analyses.

Cependant, le partage du code utilisé reste encore majoritairement volontaire. Janssen, Pritchard, and Lee (2020) soutiennent que plus d'effort et d'actions concertées doivent être mise en place afin d'améliorer l'accessibilité aux codes. Toujours selon ces auteurs, les journaux scientifiques pourraient exiger que les auteurs rendent leur code publique lors du processus de publication. D'ailleurs, les résultats d'une expérience sur les facteurs qui influencent les chercheurs à partager leur code démontre que les initiatives individuelles ne seront pas suffisantes pour une agmentation du partage du code (Krähmer, Schächtele, and Schneck

⁷une plateforme publique *code ouvert* sur laquelle nous pouvons héberger et partager notre code.

2.5 Les sciences sociales à l'ère du numérique: les enseignements de la philosophie du logiciel libre

2023). Par conséquent, rendre le code accessible devrait devenir un standard institutionnalisé.

2.4.2.3 Appropriation capitaliste

Dans ce cas-ci, il s'agit plutôt d'un défi auquel le logiciel libre est confronté plutôt qu'une critique quant aux limites de son utilisation. En fait, l'accès au code source ainsi que la liberté et la possibilité de contribuer au développement du logiciel constitue un avantage intéressant pour les compagnies privées. Par conséquent, nous avons assisté à une intégration partielle du logiciel libre dans la logique capitaliste (Broca 2013; Bessen 2002). Certaines d'entre elles utilisent les utilisateurs comme une main d'oeuvre gratuite afin de bonifier leur logiciel, ce qui permet, dans certains cas, de générer des revenus commerciaux dont l'entreprise est la seule bénéficiaire (Couture 2020). Attention, il ne faut pas penser que toutes les compagnies agissent de manière prédatrice. Le but ici est de souligner que certaines pratiques commerciales trouble l'essence du mouvement du logiciel libre, qui se veut davantage être un outil de collaboration accessible, plutôt qu'un moyen pour générer des profits. Il est important de garder en tête les valeurs et la philosophie qui a donné lieu à ce mouvement.

2.5 Les sciences sociales à l'ère du numérique: les enseignements de la philosophie du logiciel libre

Ce chapitre a voulu mettre de l'avant le logiciel libre afin d'initier les lecteurs et lectrices à ce monde. Le but n'était pas de présenter de manière exhaustive tout ce champ. Plutôt, nous avons préféré nous limiter aux bases de compréhension, ainsi qu'à quelques exemples. Par conséquent, nous souhaitons qu'à la lecture du chapitre, les lecteurs et lectrices soient mieux outillés pour comprendre et réfléchir par rapport à ce monde, et ainsi insérer ces réflexions dans leurs démarches scientifiques. Générer

2 Le monde du libre

des idées et des débats nous paraît bien plus promoteur pour l’avenir que d’apprendre par coeur.

En guise de conclusion, nous souhaitons résumer ce chapitre tout en situant ces différents éléments dans les sciences sociales à l’ère du numérique. Le livre de Marres (2017) est très intéressant à ce sujet. Face au constat que la vie sociale se trouve affectée par les changements numériques, il nous faut en tant que chercheur du monde social réfléchir à notre façon de comprendre les changements qui sont entrain de s’opérer. Bien que ces réflexions ratissent large ⁸, nous nous concentrons ici sur la dimension méthodologique.

Comme nous l’avons présenté ci-haut, les bas coûts associés à l’utilisation ainsi que la facilité du partage avec la communauté nous semble être deux avantages importants pour l’avenir des sciences sociales numériques. Notamment parce qu’ils ont le potentiel d’améliorer la transparence des protocoles scientifiques. Dans *Designing Social Inquiry*, l’un des livres les plus influents en science politique depuis les trente dernières années, les auteurs définissent quatre caractéristiques que chaque recherche doit posséder afin d’être considérée comme scientifique (King, Keohane, and Verba 2021a). L’une d’elles, est que la *procédure doit être publique*: “La recherche scientifique utilise des méthodes explicites, codifiées et publiques afin de générer et analyser des données sur lesquelles la fiabilité peut ensuite être déterminer” (King, Keohane, and Verba 2021a, 6). Chaque individu qui souhaite contribuer à la connaissance et à la compréhension globale que nous avons de la réalité sociale doit garder en tête cette caractéristique fondamentale. Comme nous l’avons exposé, le partage du code devient un impératif pour assurer la transparence, la répliquabilité ainsi que la qualité des recherches.

⁸Allant de nos postulats ontologiques, épistémologiques et méthodologiques.

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

3.3 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.

Quelques méthodes sont couramment utilisées pour la collecte de données : le sondage, le panel, l'enquête, ou le questionnaire. Cette section se concentre sur le sondage. 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,

3 Les outils de collecte de données

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, 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.

3.3.1 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: LimeSurvey, Zoho Survey, Qualaroo, Formstack, Wufoo, Checkbox Survey, SmartSurvey, QuickTap-Survey, SoGoSurvey, Snap Surveys, AskNicely, Opinio, Alchemer, Cognito Forms, Feedbackify.

3.3.1.1 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

3.3 Plateformes de sondages et collecte de données

sondages. Par exemple, Qualtrics s'adapte à différents formats en fonction de l'appareil du répondant (Evans & Mathur, 2018).

3.3.1.2 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.

3.3.1.3 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.

3.3.1.4 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.

3.3.1.5 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).

3.3.2 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

3.3 Plateformes de sondages et collecte de données

éthiques et de confidentialité (Nayak & K. A., 2019). Comme la généralisation est 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 valides, 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 vous référer directement à cet 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

3.4 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. Il vous est donc recommandé de vous renseigner davantage avec d'autres ressources afin de compléter ce qui est indiqué dans cet ouvrage. “

3.4 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 : immigrat*. 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. La personne chercheuse pourrait, par exemple, rechercher des articles sur

3.4 Factiva : outils de récolte de données médiatiques

les immigrants syriens, et rajoutant les opérateurs “and” ou encore “or”, de même que le mot « syri* », l’étoile étant rajoutée pour inclure le plus de mots possible.

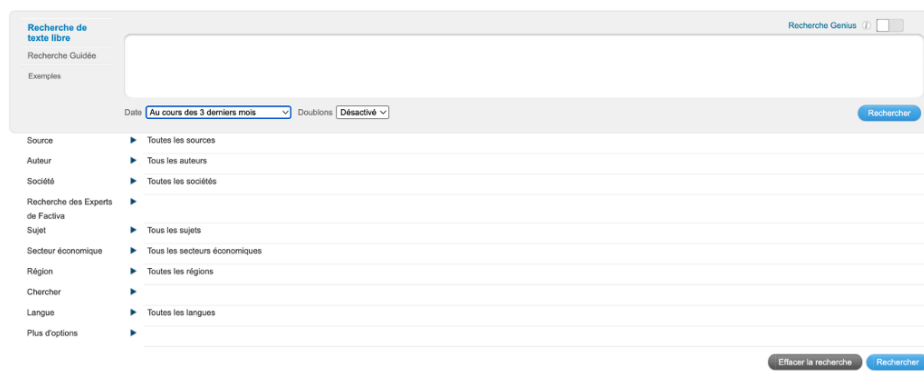
The image shows the Factiva search interface. At the top, there's a search bar with the text "Recherche de texte libre" and a "Recherche Genius" toggle. Below the search bar, there's a "Date" dropdown menu set to "Au cours des 3 derniers mois", a "Doublons" dropdown set to "Désactivé", and a "Rechercher" button. On the left side, there's a sidebar with various filters: "Source", "Auteur", "Société", "Recherche des Experts de Factiva", "Sujet", "Secteur économique", "Région", "Chercher", "Langue", and "Plus d'options". Each filter has a dropdown arrow and a list of options. At the bottom right, there are two buttons: "Effacer la recherche" and "Rechercher".

Figure 3.1: image3_1

Ainsi, Factiva permet d’avoir accès facilement à des données utiles pour de l’analyse textuelle d’articles médiatiques. Comme les textes deviennent accessibles aux chercheurs.euses, ils permettent de faire facilement une analyse de contenu par thèmes ou par ton.

Cependant, ce ne sont pas tous les médias qui sont accessibles sur Factiva. Dans l’optique où un média recherché n’est pas trouvable sur Factiva, le logiciel Eureka représente une bonne alternative. Eureka se concentre principalement sur les médias francophones (autant au Québec qu’en Europe). La structure d’Eureka est similaire à celle de Factiva. En effet, Eureka permet de filtrer des articles médiatiques par requête de recherche adaptée à la source, la date ou encore l’auteur. Toutefois, les requêtes de recherche doivent être formulées d’une manière quelque peu différente. Elles doivent donc être adaptées au fonctionnement d’Eureka. Les articles doivent être sélectionnés à la main, et peuvent être téléchargés dans un document PDF pouvant contenir un maximum de 50 articles à la fois. La Figure 2 contient l’interface de recherche d’Eureka.

3 Les outils de collecte de données

The screenshot displays the EUREKA search platform's advanced search interface. At the top, a navigation bar includes the EUREKA logo, a 'RECHERCHER' button, and links for 'DOSSIERS' and 'PUBLICATIONS PDF'. The main heading is '< Recherche avancée'. Below this, there's a section for 'Mots clés dans tout le texte' with a large text input field. To the right of this field are three rows of search filters, each with 'ET', 'OU', and 'SANS' options and a dropdown menu for 'dans le titre', 'dans l'introduction', and 'dans le nom de l'auteur'. A button 'Ajouter une zone de mots clés' is positioned below these filters. The 'Sources' section allows users to select sources by 'groupe de sources', 'critères de sources', or 'nom de source'. Below this is a 'Domaine de recherche' dropdown menu set to 'Tout le contenu'. A 'Période' dropdown menu is set to 'Depuis 30 jours'. At the bottom, there are 'Recommencer' and 'Recherche' buttons. On the right side, a 'Astuces de recherche' sidebar provides examples of search queries and their results, such as 'pomme verte' and 'bières blondes'.

Figure 3.2: image3__2

Il existe aussi une panoplie d'outils permettant un accès à des données médiatiques. Quoique Factiva soit intuitive et que de nombreuses universités possèdent des licences permettant d'exploiter la plateforme, plusieurs alternatives existent pour les personnes chercheuses. NexisUni, qui comprend entre autres l'outil LexisNexis Academic particulièrement prisé par le champ d'études de communication aux États-Unis, représente une excellente alternative. C'est également le cas de NewsBank qui permet lui aussi un accès à un vaste répertoire d'articles médiatiques. Les chercheurs.euses peuvent choisir la plateforme qui leur convient le mieux, en prenant en compte notamment l'accès qui peut leur être fourni par l'institution universitaire les employant.

En somme, la révolution numérique permet un accès sans précédent aux données médiatiques, ce qui permet des analyses approfondies du rôle des médias traditionnels dans une société démocratique.

3.5 Les extracteurs : avoir accès à des données massives via du code.

3.5 Les extracteurs : avoir accès à des données massives via du code.

Chacun des acteurs démocratiques énumérés précédemment peut également être étudié par le biais d'extracteurs qui offrent un accès à des données numériques massives. Les extracteurs de données numériques sont des infrastructures de code permettant d'extraire des données brutes d'une source définie. La section suivante explique comment les extracteurs peuvent être utiles dans un contexte de recherche en sciences sociales numériques.

Les données en lien avec les décideurs sont souvent accessibles sur des sites gouvernementaux. Toutefois, certaines identifications peuvent être nécessaires et l'accès peut être compliqué, particulièrement dans une perspective de données massives. C'est dans cette optique que les extracteurs de données numériques peuvent être utiles. Un code peut extraire de manière automatisée les débats des parlements, les communiqués de presse des gouvernants, les plateformes électorales des partis politiques, ce qui offre un accès inégalé aux chercheurs.euses aux données de décideurs. Dans une autre optique, des extracteurs peuvent également offrir l'accès aux données provenant de médias socionumériques comme Twitter (maintenant X) ou Facebook . Un extracteur peut, par exemple, être en mesure de répertorier l'ensemble des Tweets de journalistes, de politiciens ou encore de citoyens de manière automatisée, offrant un accès inégalé aux chercheurs.euses à des données massives exclusives. L'élaboration d'extracteurs est toutefois facilitée par l'existence d'API (Application programming interface) sur les plateformes exploitées. L'API d'un site ou d'une application permet à un tiers parti d'avoir accès à du code expliquant le fonctionnement de la plateforme étudiée, ce qui en facilite l'extraction de données. Par exemple, Twitter possédait avant les changements de directions récents un API qui facilite l'élaboration d'un extracteur. En contrepartie, Facebook ne possède pas d'API, ce qui rend l'accès à ses données beaucoup plus complexe. Un extracteur peut également offrir l'accès à des données médiatiques, en

codant un accès à des fils RSS ou encore aux HTML des médias extraits.

L'élaboration d'un extracteur est toutefois une tâche complexe qui requiert un certain nombre de connaissances en lien avec les langages de programmation. Les chapitres 4 et 5 du présent ouvrage offrent justement un survol du langage fonctionnel R, qui est utilisé par de nombreux développeurs lors de l'écriture d'extracteurs. R est également reconnu pour ces fonctionnalités statistiques qui sont, elles aussi, abordées ultérieurement dans ce livre.

3.6 Covidence : outil de récolte d'articles scientifiques

Comme mentionné précédemment, les outils numériques de données massives facilitent le travail des personnes chercheuses lors de la récolte de données dans le cadre d'analyses empiriques. Cependant, la révolution technologique offre également des outils pouvant être utiles lors d'autres étapes du cycle de la recherche. Il s'agit notamment du cas de la revue de littérature, alors que de nombreux outils offrent aux personnes chercheuses des ressources permettant d'élaborer un cadre théorique exhaustif par le biais de données massives sur la littérature scientifique. L'outil Covidence, géré par une compagnie sans but lucratif, en est un exemple particulièrement prisé du monde académique lors de l'entreprise de revues de littérature.

La plateforme en ligne Covidence est utilisée pour faciliter les revues systématiques de littérature. Cette dernière permet de réduire drastiquement le temps d'accomplissement du travail en plus de le rendre plus simple et plus intuitif. L'outil a été développé pour mieux gérer et organiser l'évaluation de quantité importante d'études scientifiques. L'exécution d'une revue de littérature sur Covidence se fait par le biais d'un double codage. C'est-à-dire que l'évaluation des études se fait manuellement par deux codeurs

3.6 Covidence : outil de récolte d'articles scientifiques

travaillant de manière autonome et qui mettront en commun leurs résultats à la fin de l'exercice. L'outil est reconnu pour ses trois étapes précises : « Title and abstract screening », « Full text review » et « Extraction ». Covidence permet d'importer des données massives provenant de base de données bibliographiques. En effet, l'outil lance des requêtes auprès de multiples bibliothèques, ce qui offre l'accès à des milliers d'études sur le champ étudié par les personnes chercheuses. Ces requêtes sont adaptées aux besoins spécifiques de la personne chercheuse voulant explorer en profondeur un domaine de la littérature scientifique.

La première étape, soit le « Title and abstract screening », consiste en la révision des titres et des résumés des articles récoltés. Pour rendre le travail davantage efficace, il est nécessaire d'inclure des critères précis pour analyser les titres et résumés d'articles. En se servant du jugement et des critères qui étaient recherchés, les individus doivent éliminer ou accepter selon la pertinence de l'article quant à la littérature étudiée. Cette partie est souvent longue, puisque la littérature existante est souvent massive. Il est donc important pour les personnes chercheuses de se rencontrer à maintes reprises pour discuter des conflits de jugement et pour trouver des compromis. En outre, cette étape, plutôt longue, s'avère très utile et motivante, puisqu'il est possible de développer un jugement critique davantage raffiné et de s'instruire dans une littérature continuellement plus précise.

Une fois avoir complété la revue des titres et des résumés, il faut entamer le « Full text review » qui, comme l'indique le nom, consiste à la révision complète des textes sélectionnés. Cette étape demande d'analyser chaque texte, puis de voter « oui », « non » ou « peut-être » quant à la conservation du texte dans la revue de littérature. Le vote permet donc soit d'exclure l'article, de le retenir ou de l'envoyer à la prochaine étape. D'un autre côté, les conflits rendent le travail beaucoup plus long, puisque les codeurs.euses ont un texte entier à argumenter. Ainsi, cette partie du travail, bien qu'elle comporte beaucoup moins de documents, est assez longue et exigeante.

La dernière étape, soit celle de l'extraction, consiste à recueillir toute don-

3 Les outils de collecte de données

née étant utile à l'étude de la littérature désignée. Cette étape est demandante, car les chercheur.euse.s doivent se conformer à une grille de codification prédéfinie. Le but est qu'un consensus entre les codeurs émerge de ce processus. L'extraction permet de faire ressortir les théories, les méthodologies et les conclusions présentent dans les études retenues.

Une fois les étapes de la revue systématique terminées, Covidence facilite l'exportation des résultats de l'extraction sous forme de tableaux, de graphiques et de rapports pour la méta-analyse ou pour la rédaction d'articles scientifiques. De nombreuses universités offrent un accès à Covidence par le biais de licences, et l'outil est particulièrement utile et bien construit. Toutefois, il existe d'autres alternatives à Covidence. Le choix de l'outil dépend des coûts de même que des besoins spécifiques des personnes chercheuses. Les plateformes DistillerSR, Archie et Rayyan sont notamment largement utilisées par les personnes chercheuses.

3.7 Conclusion et discussion:

Le précédent chapitre portait sur les différents outils de collecte de données massives mis à la disposition des chercheur.euse.s s'intéressant au champ des sciences sociales numériques. Les outils relevés se démarquent par leur capacité d'accorder l'accès à des données permettant d'étudier les trois principaux acteurs de la société démocratique, soit: les citoyens, les décideurs et les médias. Comme mentionné à plusieurs reprises lors du chapitre, le but de ce dernier n'est pas d'offrir une liste complète des outils disponibles. Toutefois, les outils énumérés ont été sélectionnés en raison de leur intuitivité, leur relative simplicité d'accès de même que leurs capacités techniques considérées par les auteurs comme étant particulièrement pertinentes dans une optique de recherche en sciences sociales numérique. Ainsi, ce chapitre démontre que la possibilité d'effectuer des recherches en sciences sociales numériques par le biais de données massives est plus que jamais accessible à la communauté scientifique, particulièrement en ce qui a trait à la collecte de données permettant de tels travaux. Une fois les

3.7 Conclusion et discussion:

données collectées, le travail d'analyse représente un défi technique supplémentaire se dressant devant les personnes chercheuses. Les chapitres suivants visent à familiariser les chercheurs.euses à des outils méthodologiques permettant l'analyse et la visualisation de données massives au sein des sciences sociales.

Bibliographie:

Schroeder, R. (2014). Big data and the brave new world of social media research. *Big Data & Society*, 1(2), 2053951714563194.

Chadwick, A. (2017). The hybrid media system: Politics and power. Oxford University Press.

Connelly, R., Playford, C. J., Gayle, V., & Dibben, C. (2016). The role of administrative data in the big data revolution in social science research. *Social science research*, 59, 1-12.

Manovich, L. (2011). Trending: The promises and the challenges of big social data. *Debates in the digital humanities*, 2(1), 460-475.

Burrows, R., & Savage, M. (2014). After the crisis? Big Data and the methodological challenges of empirical sociology. *Big data & society*, 1(1), 2053951714540280.

Kramer, A. D., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National academy of Sciences of the United States of America*, 111(24), 8788.

Andrade, C. (2020). The Limitations of Online Surveys. *Indian Journal of Psychological Medicine*, 42(6), 575-576. <https://doi.org/10.1177/0253717620957496>

Evans, J. R., & Mathur, A. (2018). The value of online surveys: A look back and a look ahead. *Internet Research*, 28(4), 854-887. <https://doi.org/10.1108/IntR-03-2018-0089>

3 Les outils de collecte de données

Nayak, M., & K A, N. (2019). Strengths and Weakness of Online Surveys. 24, 31-38. <https://doi.org/10.9790/0837-2405053138>

4 R ou ne pas R?

Plusieurs notions liées à l'ère numérique, notamment à ce qui a trait aux opportunités et difficultés que cette dernière peut amener, ont été présentées par l'entremise du chapitre précédent. C'est un monde de possibilité qui s'offre à ceux qui maîtrisent les nouveaux outils des temps modernes. Mais comment en arriver là ? Le présent chapitre a pour but de présenter certains outils flexibles et péreins permettant la réalisation de nombreuses tâches. Une des premières étapes permettant de notamment réaliser la collecte, l'analyse et la visualisation graphique de données ainsi que la rédaction de documents est l'apprentissage d'un langage de programmation. Bien que plusieurs langages de programmation existent, le présent ouvrage priorise le langage **R**. Les sections suivantes présentent ce langage de programmation, ces forces et ces faiblesses ainsi que les raisons de son utilisation. Enfin, la dernière section présente un environnement de programmation qui se prête bien à son utilisation.

4.1 Pourquoi R?

Comme mentionné précédemment, il existe plusieurs langages de programmation. **R** a deux types de compétiteurs : les logiciels à licences comme SAS, STATA et SPSS, et les langages *OpenSource* tels que Python et Julia. **R** est un langage de programmation *OpenSource* développé par des statisticiens, pour des statisticiens, dans les années 1990 (Tippmann 2015). **R** prend ses racines dans le langage de programmation S, créé notamment par Ross Ihaka et Robert Gentleman. Ces derniers ont fait des choix non

4 *R* ou ne pas *R*?

orthodoxes lors de l'élaboration du langage, qui font aujourd'hui la popularité de ce logiciel auprès d'un large pan de la communauté académique. En effet, Morandat et al. (2012) rapporte que le langage a été élaboré afin qu'il soit intuitif et qu'il permette aux nouveaux utilisateurs de rapidement réaliser des analyses.

Le langage de programmation **R** a plusieurs avantages qui font de lui un outil puissant et utile pour tout chercheur. L'un de ses grands avantages est qu'il est *OpenSource*. Ayant déjà abordé le sujet dans le chapitre précédent, il sera question ici de simplement rappeler les grandes lignes de l'argument, à savoir que : 1) l'*OpenSource* est gratuit d'utilisation; 2) l'*OpenSource* est développé de façon bottom-up, ce qui lui procure une grande flexibilité; et 3) il permet aux utilisateurs de créer leurs propres fonctions. À l'inverse, les logiciels à licences sont coûteux, rigides et l'ajout de fonctionnalités se fait par les développeurs internes à la compagnie. Ces formalités rendent le processus plus lent et réduisent l'éventail des possibilités pour la personne chercheuse. Ceci étant dit, certains avanceront que c'est justement ce processus interne lent qui assure la validité et la fiabilité des analyses effectuées par SAS, STATA ou SPSS. Or, dans son livre dédié aux utilisateurs de SPSS et de SAS, Muenchen (2011) soulève le point que bien souvent, ce sont des individus atomisés qui développent les nouvelles fonctionnalités de ces langages et que le processus de révisions se fait ensuite par des comités internes de testeurs. Il en va de même pour le développement des *packages* R dans la mesure où ce dernier se voit testé et amendé par plusieurs programmeurs indépendants dans un processus itératif des plateformes telles que GitHub. De plus, bien des nouvelles techniques statistiques sont développées pour R par des chercheurs qui publient leur travail dans des journaux académiques revus par des pairs, assurant la qualité du procédé. Le fait que SAS et SPSS permettent à leur utilisateur d'intégrer des routines R à leur programme est un indicateur fort ne serait-ce que de l'utilité de R (Muenchen 2011). Le langage de programmation **R** permet également de réaliser une grande quantité de tâches de recherche. En effet, les personnes programmant en **R** peuvent notamment manipuler et visualiser des données, faire différents types d'analyses,

4.2 Où coder en R ?

créer des fonctions et faire des boucles en plus de pouvoir combiner **R** avec certains langages de balisages.

D'un autre côté, l'utilisation du langage de programmation **R** peut être perçue comme ayant certains inconvénients. Plusieurs disent que la courbe d'apprentissage peut être plus grande que celle de programmes à licences. La véracité de cet argument est discutable. Les programmes demandant des licences ont également un coût d'entrée. De plus, les nouvelles itérations de ces logiciels amènent des changements demandant une période d'adaptation pour la personne chercheuse. D'autres disent que le développement *OpenSource*, spécifiquement celui du langage de programmation **R**, se fait de façon anarchique. Cela est davantage une question d'opinion et de conception du monde qu'une vérité. Le développement de *package* se fait effectivement de manière décentralisée et toute personne sachant programmer en **R** peut collaborer à cette communauté. Bien qu'il n'y ait pas d'autorité centrale, les *packages* sont regroupés sur le *Comprehensive R Archive Network* (CRAN) (voir le <https://cran.r-project.org/> pour plus d'information). Le site a une politique de dépôt stricte, ainsi les *packages* doivent être suffisamment documentés. Il est également possible d'y télécharger le langage de programmation **R**. Ce langage, ainsi que ces différents *packages*, sont disponibles sur Windows, macOS et Linux.

4.2 Où coder en R ?

Un environnement de développement intégré (IDE) permet aux programmeurs de consolider les différents aspects de l'écriture d'un programme informatique. Ils permettent de réaliser toutes les activités courantes d'un programmeur – l'édition du code, la construction des exécutables et le débogage – au même endroit. Les environnements de développement intégrés sont conçus pour maximiser la productivité du programmeur. Ils fournissent de nombreuses fonctionnalités – notamment la coloration syntaxique ainsi que le contrôle de version – pour créer, modifier et compiler du code. Certains environnements de développement intégré sont dédiés

4 *R* ou ne pas *R*?

à un langage de programmation spécifique. Par conséquent, ils contiennent des fonctionnalités qui sont plus compatibles avec les paradigmes de programmation du langage auquel ils sont associés. Enfin, il existe de nombreux environnements de développement intégré multilingues.

Comme mentionné précédemment, *R* est un des langages de statistiques et d'exploration de données les plus populaires en sciences sociales. *R* est pris en charge par de nombreux environnements de programmation. Plusieurs ont été spécialement conçus pour la programmation en *R* – le plus notable étant RStudio – tandis que d'autres sont des environnements de programmation universels – tels que Visual Studio Code – et prennent en charge *R* via des plugins. Il est également possible de coder en *R* à partir d'une interface en ligne de commande. Une telle méthode permet la communication entre l'utilisateur et son ordinateur. Cette communication s'effectue en mode texte : l'utilisateur tape une « ligne de commande » – c'est-à-dire du texte dans le *terminal* – pour demander à son ordinateur d'effectuer une opération précise, telle que rouler un fichier de code *R*.

La suite du chapitre présente RStudio, notamment à travers ses avantages et inconvénients, mais également des exemples de ses fonctionnalités ainsi que des conseils sur comment l'utiliser et le personnaliser.

4.3 Qu'est-ce que RStudio ?

RStudio est un projet open source destiné à combiner les différentes composantes du langage de programmation *R* en un seul outil (Allaire, 2011). RStudio fonctionne sur tous les systèmes d'exploitation, y compris Windows, Mac OS et Linux. En plus de l'application de bureau, RStudio peut être déployé en tant que serveur pour permettre l'accès Web aux sessions *R* s'exécutant sur des systèmes distants (Allaire, 2011). RStudio facilite l'utilisation du langage de programmation *R* en offrant de nombreux outils permettant à son utilisateur d'aisément réaliser ses tâches. Parmi les plus utiles, on retrouve notamment une fenêtre d'aide, de la

4.3 Qu'est-ce que RStudio ?

documentation sur les différents packages R, un navigateur d'espace de travail, une visionneuse de données et une prise en charge de la coloration syntaxique (Horton, Kleinman, 2015). De plus, RStudio permet de coder dans plusieurs langages et de supporter une grande quantité de formats. Il fournit également un support pour plusieurs projets ainsi qu'une interface pour utiliser des systèmes de contrôle, tels que GitHub (Horton, Kleinman, 2015).

RStudio a plusieurs avantages. Son utilisation est facile à apprendre pour les débutants. Les principaux éléments d'un IDE sont intégrés dans une disposition à quatre volets (Verzani, 2011). Cette disposition comprend une console, un éditeur de code source à onglets pour organiser les fichiers d'un projet, un espace pour l'environnement de travail et un quatrième volet où il est notamment possible d'afficher des graphiques ou de la documentation sur différents packages. Ce volet permet d'ailleurs d'accéder au répertoire des *packages* disponibles pour R en plus de permettre à l'utilisateur de consulter l'arborescence de ses fichiers. De plus, on y retrouve la possibilité de créer plusieurs espaces de travail – appelés projets – qui facilitent l'organisation de différents *workflows*.

Il y a plusieurs autres aspects de RStudio que les programmeurs apprécient. Parmi ceux-ci se trouve le fait qu'il peut être utilisé via un navigateur Web pour un accès à distance (Verzani, 2011). De plus, RStudio supporte plusieurs langages de programmation ainsi que différents langages de balisage. Qui plus est, de nouvelles fonctionnalités sont régulièrement ajoutées pour satisfaire les besoins de la communauté scientifique. Enfin, R logiciel est également souvent mis à jour.

Parmi ce que certains considèrent comme étant les points faibles de RStudio, on retrouve des éléments liés à la configuration. Certains utilisateurs trouvent que le nombre de raccourcis est limité. D'autres trouvent que le *set up* des différents panneaux n'est pas ergonomique, ou même qu'il n'est pas possible de pouvoir suffisamment personnaliser l'environnement de programmation. De plus, certains utilisateurs ont rapporté que RStudio

4 R ou ne pas R?

était plus lent que d'autres alternatives pour quelques opérations, surtout celles comprenant de longs codes.

4.4 Comment utiliser RStudio ?

Bien que de nombreux éléments puissent être personnalisés, la disposition par défaut de RStudio est composée de quatre volets principaux (Verzani, 2011). Dans le coin supérieur gauche se trouve le cadran principal. C'est dans celui-ci que l'utilisateur passera la plus grande partie de son temps. On y modifie des fichiers de différents formats et il est possible d'y afficher des bases de données. Dans le coin inférieur gauche se trouve la console ainsi que le terminal. Dans cette première, on peut interagir avec R de la même manière que dans le cadran principal, mais le code ne sera pas enregistré. Le terminal, pour sa part, est le point d'accès de communication entre un usager et son ordinateur. Bien que les différents systèmes d'exploitation viennent avec un terminal déjà intégré, il est aussi possible d'y accéder à partir de RStudio.

On retrouve, dans le coin supérieur droit, l'espace de travail. Ce cadran contient trois éléments : *l'environnement global*, *l'historique* et *les connections*. *L'environnement global* est l'endroit où l'utilisateur peut voir les bases de données, les fonctions et les différents autres objets R qui sont actifs. Il peut cliquer sur les divers éléments actifs pour les consulter. L'onglet *historique* permet à l'utilisateur de consulter les derniers morceaux de code R qu'il a roulé ainsi que les dernières commandes écrites dans la console. L'onglet *connections*, pour sa part, permet de connecter son IDE à une variété de sources de données et d'explorer les objets et les données qui la composent. Il est conçu pour fonctionner avec une variété d'autres outils pour travailler avec des bases de données en R dans RStudio.

Le cadran dans le coin inférieur droit, pour sa part, contient plusieurs outils très utiles pour les usagers de RStudio. L'onglet *Files* permet à l'utilisateur de naviguer dans les fichiers que contient son ordinateur

4.5 Personnaliser son RStudio

sans avoir à sortir de RStudio. L'onglet *Plots* permet de visualiser les graphiques générés à partir de R, que ce soit en utilisant *ggplot2*, *lattice* ou *base R*. L'onglet *Packages* permet de consulter les packages installés précédemment par l'utilisateur en plus de pouvoir en consulter la documentation. C'est aussi un des différents endroits à partir d'où il est possible d'installer des packages avec RStudio. L'onglet *Help* permet à l'utilisateur de chercher et de consulter de la documentation sur de nombreux sujets, notamment sur les différentes fonctions en R ainsi que sur les packages. Pour sa part, l'onglet *Viewer* permet la visualisation de contenu web local.

Enfin, l'utilisateur peut modifier les dimensions par défaut pour chacun des quatre cadrans principaux. En cliquant sur la division des sections, il est possible d'ajuster l'allocation horizontale de l'espace. De plus, chaque côté dispose d'un autre séparateur pour ajuster l'espace vertical. Qui plus est, la barre de titre de chaque cadran comporte des icônes pour ombrer un composant, maximiser un cadran verticalement ou modifier la taille de l'espace de travail (Verzani, 2011; Nierhoff et Hillebrand, 2015).

4.5 Personnaliser son RStudio

5 Baliser les sciences sociales : langages et pratiques

Lorsque vous lisez une page Web, un article scientifique ou un curriculum vitae professionnel, vous vous doutez peut-être que le texte n'est pas toujours produit à l'aide d'un logiciel de traitement de texte comme Microsoft Word, Apple Pages ou LibreOffice Writer. La mise en page complexe réglée au millimètre près, la qualité des figures et des tableaux, l'utilisation de gabarits professionnels, le style des références ou encore la présence d'éléments interactifs sont difficiles et parfois impossibles à reproduire à l'aide d'un logiciel de traitement de texte régulier. L'ajout d'extraits de code, de tableaux de régression ou encore de figures de haute qualité graphique, ainsi que leur personnalisation, nécessitent une interface particulière.

Pour ces raisons et plusieurs autres, les chercheurs en sciences sociales font souvent appel aux langages de balisage, ou *markup languages*. Ceux-ci permettent de produire des documents et pages Web sans les limitations des logiciels de traitement de texte. Le présent livre, par exemple, est écrit à l'aide du langage de balisage Markdown avec l'aide du système de publication Quarto. D'entrée de jeu, vous vous demandez peut-être quelle est l'utilité d'apprendre ces langages alors que les logiciels de traitement de texte sont nombreux, simples d'approche et en amélioration constante. Ce chapitre n'a pas pour objectif de décourager l'utilisation de ces logiciels, qui sont utiles et même souvent essentiels pour la production rapide de documents ainsi que pour des tâches de suivi des modifications et de travail avec des équipes multidisciplinaires. Le chapitre tentera plutôt de

répondre, tour à tour, aux trois grandes questions suivantes : *Qu'est-ce qu'un langage de balisage? Quand et pourquoi utiliser un langage de balisage? Comment utiliser un langage de balisage?* L'accent sera mis sur Quarto ainsi que sur les langages Markdown et \LaTeX , bien que d'autres langages soient aussi abordés.

5.1 Qu'est-ce qu'un langage de balisage?

Un langage de balisage constitue un ensemble de commandes qui peuvent être entremêlées à du texte afin de produire une action informatique. Chaque langage contient son propre ensemble de commandes cohérentes et complémentaires. De manière plus formelle, ces commandes sont nommées *balises* (*tags* en anglais) et inscrites par le chercheur ou la chercheuse au travers du texte. Les balises constituent une manière de communiquer avec le logiciel utilisé dans un langage qu'il peut comprendre. Par exemple, une balise permet d'indiquer au logiciel que vous désirez qu'une section du texte soit écrite en caractères gras, en italique, à double interligne ou encore que vous souhaitez positionner une image d'une certaine manière au travers du texte. Cette interaction est rendue possible par la standardisation des langages de balisage : chaque balise correspond à une action précise, peu importe le logiciel utilisé, la langue dans laquelle le texte est rédigé, le type d'ordinateur utilisé, etc. Dans votre document source, les balises sont entremêlées au contenu de votre document. Au moment de compiler ce dernier, les balises produisent les actions informatisées qu'elles commandent et laissent comme document final le contenu mis en page tel que vous l'avez défini via les balises utilisées. La compilation est le processus par lequel un document écrit en langage de balisage est transformé en fichier textuel, en format PDF dans le cas de \LaTeX par exemple.

Le premier langage de balisage, le Generalized Markup Language (GML), a été inventé en 1969 par les chercheurs Charles F. Goldfarb, Ed Mosher et Ray Lorie pour la compagnie IBM. Goldfarb et ses collègues devaient intégrer trois applications créées avec des langages différents et avec une

5.1 Qu'est-ce qu'un langage de balisage?

logique différente pour les besoins d'un bureau de droit. Même après avoir créé un programme qui permettait aux trois applications d'interagir, ces langages demeuraient différents et avaient chacun leur propre fonctionnement. Le développement de GML a permis de résoudre ce problème en standardisant et en structurant le langage : les mêmes commandes étaient utilisées pour accomplir les mêmes tâches dans chaque programme (Goldfarb 1996). GML a été amélioré durant les décennies suivantes et a été suivi par d'autres langages de balisage, dont \LaTeX (1985), \BibTeX (1988), HTML (1993), XML (1998), Markdown (2004) et R Markdown (2012) (Encyclopaedia Britannica 2023; Hameed 2023; Markdown Guide 2023; World Wide Web Consortium (W3C) 1998; Xie 2023).

Les langages de balisage permettent d'effectuer différentes tâches. HTML, qui est sans doute le plus connu des langages de balisage, permet de formater des sites Web. XML, quant à lui, permet de structurer de larges volumes de données. \LaTeX permet pour sa part de formater du texte et de créer des documents en format PDF. Markdown permet également de créer des documents en format PDF, mais aussi en format HTML ou DOCX — format utilisé pour les documents Word —, contrairement à \LaTeX . R Markdown permet d'ajouter des extraits de code R à un fichier en langage Markdown. Enfin, depuis 2022, le système de publication scientifique et technique multilingue Quarto permet de créer des documents qui intègrent des extraits de code R, \LaTeX , Python, Julia ou JavaScript, créés dans différents types d'environnements, à un fichier en langage Markdown (Allaire 2022). \LaTeX , Markdown, R Markdown et Quarto permettent aussi d'intégrer les références bibliographiques du système de traitement de références \BibTeX . Les langages de balisage communiquent ainsi souvent les uns avec les autres au sein d'un même fichier. Le chapitre 7 explique la manière de citer les références en langage \BibTeX par le biais de Zotero et de Better \BibTeX .

Les balises constituent une manière de donner manuellement des commandes au logiciel que vous utilisez. Si vous utilisez Microsoft Word, vous avez accès à une panoplie de boutons qui vous permettent de formater votre texte. Les balises exercent les mêmes fonctions de formatage pour