Outils de recherche en sciences sociales numériques

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

2023-08-24

Table of contents

Avant-propos

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

Introduction

1 Trois défis pour une contribution aux sciences sociales numériques

Ce premier chapitre n'est sans doute pas le plus excitant. Il ne comprend ni graphique ni exercice. Il s'ancre dans la réflexion théorique plutôt que dans la pratique méthodologique. Habituellement, c'est la partie que l'on ignore, celle que l'on saute pour passer aux « choses sérieuses ». Amateur de « choses sérieuses »? Bonne nouvelle! Cet ouvrage en est rempli. Tout comme la carrière qui s'offre à vous si vous choisissez de poursuivre dans l'étude des sciences sociales numériques.

En 2020, le monde est numérique, et rien ne semble présager un inversement de la tendance. Au contraire, celle-ci risque plutôt de s'accélérer. La pandémie de la COVID-19 a offert quelques-uns des meilleurs exemples de cette tendance: télétravail généralisé, école numérique, livraison en ligne, mobilisation via les réseaux sociaux, intelligences artificielles pour le dépistage de fausses nouvelles et application mobile pour tracer les déplacements et freiner les pandémies. L'avenir est au numérique. Pour les jeunes chercheurs en sciences sociales, cela équivaut à une montagne de «choses sérieuses».

Dans ce contexte, il ne fait aucun doute que votre carrière sera passionnante. Si vous n'en êtes pas déjà convaincu, ce livre vous fournira une panoplie d'exemples de vos nombreuses possibilités. De l'analyse textuelle dans les médias aux sondages en ligne de milliers d'individus, en passant par l'extraction de données massives des sites web ou à l'analyse de larges réseaux de communication, vous trouverez assurément des défis à la hauteur de vos aspirations.

Devant ce déluge de données numériques, le jeune chercheur peut avoir l'impression qu'il est possible, voire permis de tout faire. Entendons-nous bien: c'est presque le cas. Tous les jours, vous aurez des idées de projets plus invraisemblables les unes que les autres. Avec vos nouveaux outils, plusieurs de ces idées n'auront aucun problème à se réaliser. Le véritable problème surviendra peut-être le jour où sera négligée la réflexion théorique. La réflexion au cœur même de ce chapitre. Rappelez-vous: c'est ce chapitre que vous avez considéré sauter, au départ!

En fait, il serait surprenant que vous ne soyez pas happés, très tôt dans vos études, à des limites fondamentales à votre travail. Dans cet ouvrage, nous les appelleront « défis ». Nous ne parlons pas ici de données manquantes ou d'accès restreints à l'information. Il s'agit de défis beaucoup plus élémentaires. Ils se comptent au nombre de trois et sont à la base de toute réflexion préalable à la recherche en science sociales numériques. Ils sont:

1. Le défi technique;
2. Le défi théorique;
3. Le défi éthique.

Sachez une chose: ces défis sont présents dans toutes les grandes branches de la science, c'est-à-dire lors de la recherche, lors de la diffusion des résultats et lors de l'enseignement. Que vous comptiez opérer dans l'une, dans l'autre ou dans toutes ces branches, une bonne compréhension des trois défis permettra de limiter les risques d'impair, mais surtout d'élargir l'univers de vos possibles.

1.1 Défi #1: l'inévitable défi technique.

Le premier défi est technique, lié à l'extraction et à l'analyse des données numériques. Il nécessite l'apprentissage et le développement des méthodologies. Avec R dans sa poche, ce défi est hautement simplifié. R

permet de penser autrement les possibilités de recherche, et de travailler avec des outils tels que *Shiny*, pour la création instantanée d'applications web interactives ou *Mechanical Turk*, pour la mise en ligne de micro-tâches (*crowdsourcing*) à réaliser à faible coût par des volontaires. R facilite également la réalisation de revues de la portée de la littérature (*scoping review*), une technique permettant de cartographier la littérature scientifique dans un champ donné.

Les données massives nous entourent. Que ce soit au travers de milliers de sondages croisés, via les médias sociaux ou à l'intérieur des archives gouvernementales en ligne, il est plus facile que jamais de rassembler de grandes quantités d'information. Le défi demeure toutefois complexe lorsque vient le temps d'extraire et d'analyser ces données afin de contribuer à la connaissance scientifique.

Déjà, cet ouvrage offre une base solide sur laquelle développer vos méthodes. Celles-ci sont de plus en plus simples à apprendre et à appliquer, notamment grâce aux réseaux de collaboration en ligne. Aujourd'hui, une question peut rapidement être répondue après une recherche sur Google. *Stack Overflow* est un site Web dédié à l'entraide entre programmeurs. Vous le trouverez hautement utile.

Si les méthodologies sont plus nombreuses, efficaces et simples que jamais, beaucoup restent encore à faire pour permettre la transparence et l'accessibilité des données publiques, la collaboration entre chercheurs et l'optimisation des outils d'extractions de données. Le cœur du défi technique réside dans l'amélioration des outils utiles et nécessaires aux chercheurs.

En effet, après l'apprentissage des méthodes disponibles à l'heure actuelle, vous pourrez rapidement contribuer à leur optimisation. La beauté d'un logiciel libre comme R est qu'il vous est possible de développer de nouvelles méthodes pour faciliter la recherche, pour ensuite partager ces trouvailles avec le monde entier. Sur R, vous pourrez construire des fonctions qui accéléreront votre travail. Le développement de quatre ou cinq fonctions

pourrait ensuite faire l'objet d'un tout nouveau « package », que vous partagerez en ligne à vos pairs scientifiques.

Tous les jours, de nouveaux packages R sont développés et mis en ligne. Des dizaines existent simplement pour réaliser de l'analyse textuelle automatisée, une méthodologie qui permet l'étude quantitative de large corpus de textes. Plusieurs de ces packages, comme «Quanteda», «Topicmodels», ou ceux de la «Tidyverse» sont hautement performants, et en constante amélioration.

Il est à la portée de toute chercheuse et de tout chercheur de participer à la bonification des outils et à l'avancement des méthodologies. C'est la réponse attendu au défi technique.

1.2 Défi #2: le nécessaire défi théorique.

- Nécessite une formation selon les principaux travaux scientifiques qui étudient l'impact des données numériques sur les théories en sciences sociales:
  - On ne doit pas réinventer la roue à chaque article scientifique;
  - Comment intégrer nos travaux à la littérature actuelle?;
  - Comment faire progresser cette littérature? Démontrer l'impact des données numériques sur les théories existantes;
  - Exemple: le nationalisme: peut-on mesurer le nationalisme au travers des médias sociaux? Si oui, comment cela peut-il contribuer à la littérature sur le nationalisme?

1.3 Défi #3: l'épineux défi éthique.

- Autour des questions de l'effet de l'ère numérique sur la confidentialité, la sécurité informatique, le consentement et le droit des sujets secondaires:

- Le numérique offre beaucoup d'opportunité tout à fait légale, mais pas nécessairement éthique;
- Nécessaire d'encourager la réflexion par rapport aux défis humains entourant l'utilisation des nouvelles données numériques;
- Comment utiliser ces données pour améliorer les vies, sans brimer les libertés individuelles?;
- Exemple: intelligence artificielle (machine learning): Le milieu académique est loin d'être seul à s'intéresser à la grande quantité d'information disponible. Les partis politiques, les agences de marketing et bien d'autres organisations utilisent ces informations à des fins de victoires, ou de ventes.

1.4 Conclusion de cette première partie du chapitre

- Au travers des nouveaux apprentissages et des exemples qui sont offerts dans ce livre, le lecteur est encouragé à se poser ces 3 questions:

1. D'abord, comment puis-je utiliser ces nouveaux outils pour faire progresser les méthodologies de recherche actuelles?
2. Ensuite, comment puis-je utiliser ces nouveaux outils pour contribuer à l'avancement des théories de mes champs de recherche?
3. Enfin, comment puis-je utiliser ces nouveaux outils pour exercer un impact positif sur mes semblables?

1.5 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 commun 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 (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 dûes 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 (e.g., béhavioralisme, 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 à certains points relatifs aux impacts des données massives sur la recherche en sciences sociales. Premièrement, je présente une conceptualisation des données massives. Deuxièmement, je me penche sur les impacts des données massives en sciences sociales et souligne tout particulièrement comment elles affectent les enjeux de la *validité* interne et externe dans la

domaine des sciences sociales. Finalement, j'explore quelques pistes de réflexion sur l'avenir des données massives en sciences sociales en analysant quelques changements *épistémologiques* que ces données pourraient potentiellement entraîner.

1.6 Définition des données massives

Ce qui définie les données massives comme concept est souvent mêlé avec le phénomène social qui l'accompagne. Il est toutefois possible de demêler le tout en distinguant trois approches conceptuelles des données massives.

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

2. Deuxièmement, d'une perspective technique et technologique, les données massives constituent un (2) 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.

3. Finalement, d'une perspective sociologique, les données massives représentent (3) 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'analyses 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 vélocité et la véracité. Cette conceptualisation jouie 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 du secteur privé et public ainsi que la recherche scientifique pure (Gartner 2012).

1. Définition de base	Quantité importante de données dont la nature, le type, la source, etc. varient
2. Définition technique/technologique	Ensemble de *pratiques* 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

1.7 Les données massives et les sciences sociales

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 changement 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.

1.8 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 supposé 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. Mais 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, classifie les libertés civiles et les droits politiques des états du monde par degré, de 1 à 7, afin de construire un index allant d'autoritarisme complet à 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 classifier d'une certaine façon le phénomène et les données qui en découlent (dans le cas de *Freedom House* des codeurs indépendants).

1.9 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 pour avantage d'offrir aux chercheur.e.s 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 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 al. ont pu mesurer la censure exercée par le gouvernement chinois sur les réseaux sociaux. En utilisant des données massives nouvelles, King et al. ont donc pu observer une manifestation inédite de censure massive qui, sans de telles données, serait probablement demeurer 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).

Cependant, il faut aussi souligner que les données massives, de par 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.

Aux nombres de ces difficultés, trois questions sont particulièrement cruciales : (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.

1.9.1 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 d'un mécanisme causal clair***. 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 consiste en ce 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 (2 des 4 V) des données massives (textuelles, numériques, vidéos, etc.) peuvent miner la qualité de l'inférence causal entre un 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.

1.9.2 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 souligne 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, telle un sondage ou une série d'entrevues, traditionnellement utilisées 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. Le *Literary Digest* a effectivement prédit à tort la victoire de Alf Landon, le candidat du parti républicain, sur Franklin D. Roosevelt, le candidat démocrate, parce que l'échantillon de répondants utilisé par le *Literary*

Digest dans son sondage a surrpresenté 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 dûe 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 ainsi sous-estimé le soutien populaire de Roosevelt de plus de 20%.

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.

1.9.3 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 causalesou 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.

1.9.4 Données observationnelles

En ce qui à 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, de variable instrumentale, etc. peuvent également garantir des inférences explicatives et causales valides. Elles nécessitent toutefois plusieurs postulats plus restrictifs dont l'objectif est d'imiter ou de récréer, de la manière la plus fidèle possible, une distribution aléatoire du traitement – ce que la litérature 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 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.

1.9.5 Validité écologique et observation par sous-groupes

Les données massives peuvent aussi jouer d'autres rôles importants relatif 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 écart-types par sous-groupes étaient trop grand, rendant difficile l'estimation précise d'un paramètre comme la moyenne et impossible celle d'un coefficient –, la taille énorme des échantillons 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 approfondi de façon considérable.

	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

1.10 Vers le futur : les données massives effectueront-elles un changement dans la posture épistémologique en sciences sociales?

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 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 qui soulève au moins trois questions d'importance 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 politique en particulier)?; (3) et finalement, de manière reliée, les données massives proposent-elles ou non une manière de dépasser l'individualisme méthodologique qui caractérise les sciences sociales contemporaines?

2 Le monde du libre

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

Cette analogie illustre bien trois concepts au coeur de la philosophie de Richard Stallman, souvent considéré comme le père fondateur du logiciel libre: liberté, égalité, fraternité. Les utilisateurs de ces logiciels sont libres, égaux, et doivent s'encourager mutuellement à contribuer à la communauté. Ainsi, un logiciel libre est généralement le fruit d'une collaboration entre développeurs qui peuvent provenir des quatre coins du globe. Une réflexion éthique est au coeur du mouvement du logiciel libre, dont les militants font campagne pour la liberté des utilisateurs dès le début des années 1980. La Free Software Foundation (FSF), fondée par Richard Stallman en 1985, définit rapidement le logiciel «libre» [free] comme garant de quatre libertés fondamentales de l'utilisateur: la liberté d'utiliser le logiciel sans restrictions, la liberté de le copier, la liberté de l'étudier, puis la liberté de le modifier pour l'adapter à ses besoins puis le redistribuer [1]

[1] La redistribution doit évidemment respecter certaines conditions précises, dont l'enfreint peut mener à des condamnations

Il s'agit ainsi d'un logiciel dont le code source[2] est disponible, afin de permettre aux internautes de l'utiliser tel quel ou de le modifier à leur guise. Puisque le langage machine est difficilement lisible par l'homme et rend la compréhension du logiciel extrêmement complexe, l'accès au code source devient essentiel afin de permettre à l'utilisateur de savoir ce que le fait programme fait réellement. Seulement de cette façon, l'utilisateur peut *contrôler* le logiciel, plutôt que de se faire contrôler par ce dernier (Stallman, 1986).

2.1 Émergence et ascension

Plusieurs situent les débuts du mouvement du logiciel libre avec la création de la licence publique générale GNU[3], en 1983, à partir de laquelle va se développer une multitude de programmes libres. Depuis, la popularité des logiciels libres n'a cessé de croître, alors que des dizaines de millions d'usagers à travers le monde utilisent désormais ces logiciels. Parmi les plus populaires, on retrouve notamment le navigateur Firefox, la suite bureautique OpenOffice et l'emblématique système d'exploitation Linux, qui se développe d'ailleurs à partir de la licence GNU. Les logiciels libres ont différents usages (en passant par la conception Web, la gestion de contenu, les sytèmes d'exploitation, la bureautique…). Encore une fois, le logiciel libre est avant-tout une philosophie, voire un mouvement de société. C'est une façon de concevoir la communauté du logiciel, où le respect de la liberté de l'utilisateur est un impératif éthique central (reformuler?) (Williams et al., 2020:26). Si ce mouvement fut d'abord initié par quelques militants dans les années 1980, c'est aujourd'hui un véritable phénomène sociétal:

[http://www.softwarefreedom.org/resources/2008/shareware.html].

[2]Pour 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.

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

des milliers d'entreprises, d'organisation à but non lucratif, d'institutions ou encore de particuliers adoptent tour à tour ces logiciels, dont la culture globale et les valeurs (entraide, collaboration, partage) s'arriment avec le virage technologique de plusieurs entreprises à l'ère du numérique (retravailler, mais l'idée est là). [blablabla]

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

2.1.1 Logiciel libre et *open source*

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

2.2 Principaux avantages et inconvénients

La disponibilité du code source et le mode de développement collaboratif du logiciel libre facilitent également le transfert des connaissances et ce, au-delà des frontières. Où qu'ils soient, les institutions, les entreprises et les particuliers peuvent utiliser ces logiciels et les adapter en fonction de

leurs besoins respectifs. Par ailleurs, l'accès libre et égal de tous les internautes à l'ensemble de ces connaissances constitue un enjeu majeur pour la vitalité démocratique des sociétés à l'ère du numérique, caractérisées par une surabondace d'information.

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

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

3 Les outils de collecte de données

La révolution numérique engendrée par l'émergence du Big Data représente un important défi pour le monde des sciences sociales (Manovich, 2011; Burrows et Savage, 2014). Elle représente 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 autant les trois principaux acteurs 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 personnes chercheuses voulant exploiter le potentiel et les opportunités offertes par les données massives (Burrows et Savage, 2014). Le chapitre suivant vise à offrir un portrait de certains outils de collecte de données pouvant être exploités par les chercheur.euse.s en sciences sociales cherchant à tirer profit de la révolution numérique. Il sera, entre autres, 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 disponibles pour les personnes chercheuses visant à 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 largement 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, de 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. Au niveau des citoyens, le champ d'étude de l'opinion publique se consacre à l'analyse des comportements ou des attitudes politiques des citoyens. De plus, de nombreuses recherches visant à comprendre le rôle des citoyens en politique portent sur la l'influence de la société civile de même que sur les 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. Toutefois, la collecte de données permettant de mener à termes de telles études peut s'avérer complexe. Pour chaque pilier de la démocratie, les sections suivantes énumèrent et expliquent les capacités techniques d'outils permettant aux chercheurs 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.

3.2 Les outils de collecte de données de sondages

3.3 Factiva : outils de récolte de données médiatiques

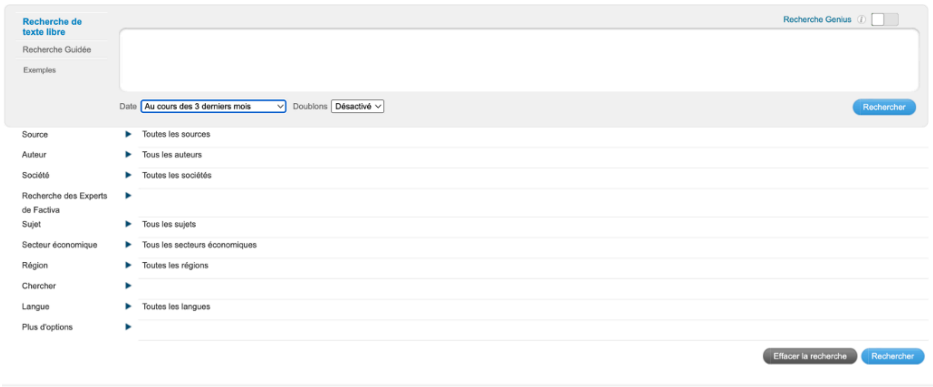
L'émergence de nouvelles technologies de même que la fragmentation médiatique causée notamment par l'apparition des chaînes de nouvelles en continu ébranlent considérablement les écosystèmes médiatiques occidentaux (Chadwick, 2017). Un récent courant de recherche se penche donc sur le rôle des médias sur le comportement des citoyens dans une perspective de fragmentation médiatique permettant aux citoyens de choisir leurs sources d'information, ce qui aurait 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 chercheur.euse.s en sciences sociales en raison de l'importante quantité de données accessibles aux personnes chercheuses, 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. Toutefois, l'accès qu'il offre aux contenus de média est particulièrement pertinent pour la communauté scientifique en communication et en sciences sociales. Il offre accès à plus de 15 000 sources médiatiques provenant de 120 pays. Il permet de télécharger une quantité illimitée de documents RTF pouvant contenir jusqu'à 100 articles médiatiques chacun. Les articles peuvent être sélectionnés automatiquement en cochant le bouton proposant de sélectionner tous les

100 articles de la page de résultat. Chaque page de résultat contient 100 articles à la fois. Factiva permet également de filtrer pour les doublons.

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

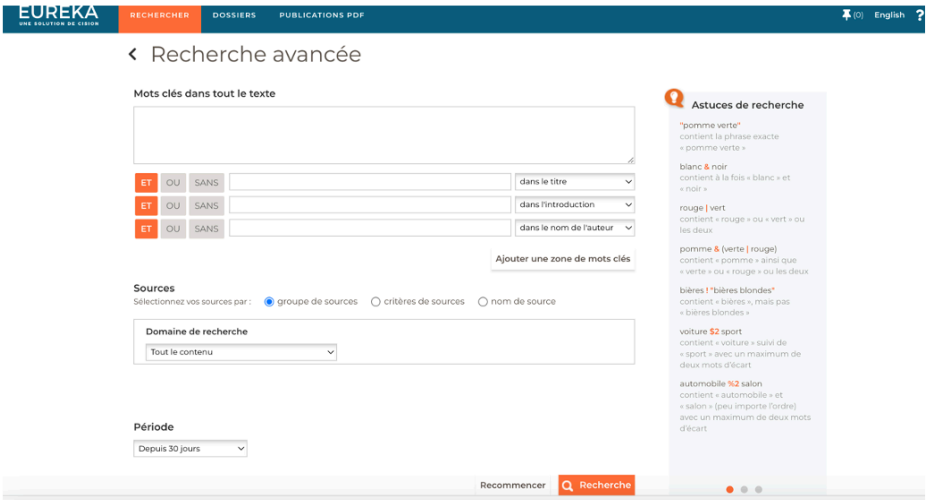
3.3.0.0.1 Figure 1. Interface de recherche de Factiva



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 deviennent facilement exploitables pour faire de l'analyse de contenu par thèmes ou par ton.

Cependant, ce ne sont pas tous les médias qui sont accessibles sur Factiva. Dans l'optique ou 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 adaptée 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.

3.3.0.0.2 Figure 2. Interface de recherche d'Eureka



Il existe toutefois une panoplie d'outils permettant un accès à des données médiatiques. Quoique Factiva soit intuitive et que de nombreuses universités possèdent des licenses permettant d'exploiter la plateforme, plusieurs alternatives existent pour les personnes chercheuses. NexisUni, qui comprend entre autre l'outil LexisNexis Academic particulièrement prisé par le champ d'étude 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 personnes chercheuses peuvent choisir la plateforme qui leur conviennent 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.

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

Chacun des acteurs démocratique énumérés précédemment peuvent également être étudiés par le biais d'extracteurs qui offrent un accès à des données massives. Les extracteurs sont des infrastructures de codes 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 numérique.

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 peuvent être utiles. Un code peut extraire de manière automatique les débats des Assemblées nationales, les communiqués de presses des gouvernants, les plateformes électorales des partis politiques, ce qui offre un accès inégalé aux personnes chercheuses aux données de décideurs. Dans une autre optique, des extracteurs peuvent également offrir accès aux données provenant des médias socionumériques comme Twitter ou Facebook, sur lesquels les acteurs politiques sont souvent très actifs. 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 encore une fois un accès inégalé aux personnes chercheuses à des données massives exclusives. L'élaboration d'extracteurs est toutefois facilitée par l'existence d'API sur les plateformes exploitées. Par exemple, Twitter possédait avant les changements de directions récents un API qui facilite l'élaboration d'un scraper. 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.

3.5 Covidence : outil de récolte d'articles scientifiques

Les outils numériques de données massives facilitent le travail des personnes chercheuses dans la récolte de données utilisées dans le cadre des 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, et cette dernière permet de réduire drastiquement le temps d'accomplissement du travail en plus de le rendre simple et 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'évalutation des études se fait manuellement par deux codeurs travaillant de manière autonome 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. 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 et une fois terminée de voter soit « 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. En revanche, avoir des conflits rend le travail beaucoup plus long puisque les codeurs.euses ont un texte entier à argumenter. Cette partie de 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 donnée étant utile à l'étude de la littérature désignée. Cette étape est demandante, car les chercheur.euse.s doivent se conformer à une grille 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 la rédaction d'articles scientifiques. De nombreuses universités offrent un accès à Covidence par le biais de license, et l'outil est patriculièrement utile et bien construit. Toutefois, 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.

3.6 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 de permettre l'accès à des données permettant d'étudier les trois principaux acteurs de la société démocratique, soit les citoyens, les décideurs et les médias. Tel que 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 pertinente dans une optique de recherche en sciences sociales numérique.

4 R ou ne pas R?

Plusieurs notions liées à l'ère numérique ainsi que certaines opportunités et difficultés que cette dernière peut amenée ont été présentées aux chapitres précédents. 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 nombreux 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, les raisons de l'utiliser ainsi qu'un environnement de programmation

4.1 Pourquoi R?

R est un langage de programmation *OpenSource* développé par des statisticiens pour des statisticiens dans les années 1990 (Tippmann 2015). C'est d'un élan d'amour propre et du désire d'honorer le langage de programmation S que Ross Ihaka et Robert Gentleman nommeront leur création, infirmant ainsi la légende selon laquelle les scientifiques seraient mauvais pour nommer les choses. Ces derniers feront des choix non orthodoxes lors de l'élaboration du langage, des choix qui font aujourd'hui la popularité de

R auprès d'un large pan de la communauté académique. En effet, Morandat et al. (2012) rapportent que le langage a été élaboré afin qu'il soit intuitif et qu'il permette aux nouveaux utilisateurs de rapidement réaliser des analyses. Ils rapportent même que dans plusieurs départements de statistiques, R est introduit en 2 semaines – environ le temps que prend l'individu moyen pour oublier ses résolutions du Nouvel An.

Toutefois, avant de débuter l'apprentissage d'un nouvel outil, il faut être convaincu de sa pertinence, de son utilité. À quoi bon apprendre à utiliser une perceuse alors que mon tournevis fonctionne parfaitement bien ? C'est pourquoi ce chapitre a deux objectifs, d'abord, il s'agira de vous convaincre de la pertinence de R suite à quoi il vous sera introduit diverses utilisations possibles de R. Plus spécifiquement, la section de réflexion théorique exposera les avantages et les inconvénients de R et le comparera à ses principaux compétiteurs. Ensuite, la réflexion méthodologique présentera brièvement la programmation de base en R et en quoi l'OpenSource fait de R un outil si puissant. Le chapitre se conclura avec quelques trucs et astuces qui vous permettront de surmonter l'anxiété que peut causer l'apprentissage d'un outil étant, pour plusieurs, quelque chose de véritablement étranger à leur relation typique avec les ordinateurs.

4.2 Réflexion théorique

R a deux types de compétiteurs lorsqu'il est question d'analyses statistiques – les logiciels à licences comme SAS,STATA et SPSS, et les langages *OpenSource*, principalement Python et sa libraire Pandas. Le chapitre précédant ayant déjà élaborer un cas exhaustif en faveur du logiciel libre, il ne sera ici que rappelé 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 ce qui rend le processus plus lent et

réduit l'éventail des possibilités. Ceci étant dit, certains avanceront que le 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 vois tester et amender par plusieurs programmeurs indépendants dans un processus itératif sur GitHub ou sur d'autres plateformes similaires. De plus, bien des nouvelles techniques statistiques sont développées pour R par des professeurs qui publie d'abord leur travail dans des journaux académiques revus par des pairs. Bien entendu, rien n'empêche un étudiant gradué de publier ses propres *packages*. C'est pourquoi Muenchen (2011) recommande de visiter le site *MACHIN* afin d'avoir une idée de la validité et de la fiabilité du *package* en question. Enfin, 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).

R n'est cependant pas qu'un outil statistique, il s'agit également d'un outil de programmation puissant. Ceci fait en sorte que le coût d'entrer de R est plus important que celui des logiciels comme SAS, STATA et SPSS puisqu'il impose l'apprentissage d'une syntaxe et d'un jargon particulier. Alors, pourquoi apprendre à programmer en plus d'apprendre à réaliser des analyses statistiques? Après tout, faire des statistique c'est déjà beaucoup! D'abord, apprendre à programmer permet de développer la résolution de problème et la logique, deux compétences aux cœur de la recherche scientifique. Programmer est au cerveau ce que courir est au coeur, il s'agit d'un exercice difficile au départ mais dont les résultats bénéfique se font sentir rapidement. Apprendre à programmer permettra également de mieux comprendre la façon dont son propre ordinateur fonctionne. Contrairement au mythe urbain qui veux que l'humain n'utilise que 10% de son cerveau, la plupart des individus ne font qu'utiliser une infime partie du potentiel de leur ordinateur. Par exemple, l'ordinateur ayant

permis aux américains d'aller sur la lune lors de la mission Apollo-11, le *Apollo Guidance Computer (AGC)*, avait 4 096 octets (bytes) de mémoire RAM[1] [*SOURCE FIABLE*]. L'ordinateur sur lequel sont écrits ces lignes a 8GB de RAM, soit approximativement 2 millions de fois plus que celle de l'AGC. Enfin, l'argument le plus probant sur la nécessité d'apprendre à programmer est celui du marché de l'emplois. Au Canada, il est prévu une pénurie de main d'oeuvre pour les emplois requérant de aptitude en statistique et en programmation comme celui de scientifique de données ou d'analyste (Employment and Social Development Canada 2023). Un rapport de PWC indique même que les employeurs devrons s'attendre à ce battre pour engager des individus compétent dans les deux domaines. Apprendre à programmer devient alors, non seulement, une façon d'améliorer votre résonnement scientifique, mais également une façon de vous démarquer sur le marché de l'emplois.

- Avantages :
  - Gratis!!
  - Plus facile d'accès que d'autres langages de programmation
  - Arguments technos :
    - Fait pour les stats
    - Fait tout ce que SPSS et Stata font sans le carcan du logiciel privé
  - Arguments d'autorités <- popularité du langage
  - Ouvre vers un monde de possibilités
  - Développement d'une expertise recherchée
- Inconvénients :
  - Courbe d'apprentissage rude pour certains

[1] La RAM (*Random Access Memory*) ou "mémoire vive" est l'espace utilisée par l'ordinateur pour enregistrer l'information directement nécessaire pour les opérations en cours d'exécution. Elle s'oppose à la ROM (*Read-Only Memory*) ou "mémoire morte" qui contient toute l'information enregistré de façon permanante dans l'ordinateur.

  - Développement anarchique
  - Risque de se perdre dans les profondeurs pleines de microbes de CRAN

- Propriété et utilisation de R
- Comparaison avec d'autres langages
  - Python, SPSS, Stata?

4.3 Réflexion méthodologique

- Base R
  - Stable mais parfois bof
  - Manipulation de données
  - Fonctions et Boucles
- La puissance de l'OpenSource
  - Peut être instable, mais souvent plus intéressant que les options en Base R
  - TidyVerse
    * Manipulation avec Dplyr
    * All hail Hadley!
  - Analyse textuelle avec ???
  - Shiny

4.4 Trucs et astuces

- 10 choses à garder en tête lorsque l'on apprend R :
  1. Vous n'allez pas briser votre ordinateur.
  2. C'est en "gossant" que l'on apprend! Essayez des trucs, expérimentez – souvenez-vous de 1.

3. Contrairement à la vraie vie, il y a toujours le crtl-z pour vous sauver.
4. Ayez de l'empathie pour vos futurs lecteurs, ou du moins pour le futur vous – COMMENTEZ VOTRE CODE!
5. Même Wozniak était mauvais au début.
6. Pensez à votre santé – levez-vous de votre chaise aux 30 minutes.
7. S'il y a un bogue, et il y en aura, *Google* est votre ami.
8. Souvent c'est une question de type de variable – caractère, numérique, facteurs, etc.
9. Si vous faites beaucoup de copier-coller de code, il y a surement une façon de l'automatiser.
10. Sérieusement, faites le 4!

4.5 Les environnements de développement intégré

4.6 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 et le contrôle de version – pour créer, modifier et compiler du code.

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

R est un des langages de statistiques et d'exploration de données les plus populaires en sciences sociales et il est open-source. Par conséquent, il est logique de choisir un environnement de programmation open-source. 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 environnement de programmation universels – tel que Visual Studio – 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 son ordinateur d'effectuer une opération précise, par exemple rouler un fichier de code R.

Le présent chapitre présente RStudio, ses avantages et inconvénients ainsi que des exemples de ses fonctionnalités de RStudio et des conseils sur comment l'utiliser et le personnaliser.

4.7 Pourquoi RStudio ?

4.7.1 Qu'est-ce que RStudio ?

4.8 Pourquoi RStudio ?

4.8.1 Qu'est-ce que RStudio ?

Comme plusieurs autres langages de programmation, R est développé grâce à des fonctions écrites par ses usagers. Un IDE, comme RStudio, est conçu pour faciliter ce travail (Verzani, 2011). RStudio est un projet open source destiné à combiner les différentes composantes du langage de programmation R en un seul outil (Allaire, 2011). Il est conçu pour

faciliter la courbe d'apprentissage des nouveaux utilisateurs. RStudio fonctionne sur toutes 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).

Figure of RStudio with some code, a plot in the bottom right corner and some data in the top right corner

RStudio facilite l'utilisation du langage de programmation R en offrant de nombreux outils permettant à son utlisateur d'aisément réaliser ses tâches. Parmi les plus utiles, on retrouve notamment une fenêtre d'aide, de la 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 supportent un grande quantité de formats. Il fournit également un support pour plusieurs projets ainsi qu'une interface pour utiliser des systèmes de contrôle des versions tels que GitHub (Horton, Kleinman, 2015).

4.8.2 Avantages et inconvénients de RStudio

RStudio a plusieurs avantages. L'utilisation de l'IDE 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 possible d'afficher des graphiques ou de la documentation sur différents packages. De plus, on y retrouve la possibilité de créer plusieurs espaces de travail – appelés projets – qui facilitent l'organisation de différents workflows.

Un autre aspect de RStudio que de nombreux programmeurs apprécient est le fait qu'il peut être utilisé via un navigateur Web pour un accès à

distance (Verzani, 2011). De plus, l'IDE offre de nombreux outils pratiques et faciles à utiliser pour gérer les packages, l'espace de travail et les fichiers. RStudio supporte plusieurs langages de programmation ainsi que différents langages de balisage. Finalement, de nouvelles fonctionnalités sont souvent ajoutées pour satisfaire aux besoins de la communauté scientifique et le logiciel est régulièrement mis à jour.

Inconvénients : - Peu de configuration, tu peux pas changer les raccourcis, etc - Très limité dans le setup des différents panneaux, tu peux pas voir 2 fichiers en même temps Tu peux pas visualiser 2 fichiers 1 à côté de l'autre, une feature de base dans n'importe quel ide Tu peux pas configurer tes raccourcis clavier, aussi une feature de base Mettons que je veux que ctrl + D copie la ligne, comme dans visual studio, je peux pas - Plus lent que d'autres alternatives pour certaines opérations

Comme plusieurs autres langages de programmation, R est développé grâce à des fonctions écrites par ses usagers. Un IDE, comme RStudio, est conçu pour faciliter ce travail (Verzani, 2011). RStudio est un projet open source destiné à combiner les différentes composantes du langage de programmation R en un seul outil (Allaire, 2011). Il est conçu pour faciliter la courbe d'apprentissage des nouveaux utilisateurs. RStudio fonctionne sur toutes 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).

Figure of RStudio with some code, a plot in the bottom right corner and some data in the top right corner

RStudio facilite l'utilisation du langage de programmation R en offrant de nombreux outils permettant à son utlisateur d'aisément réaliser ses tâches. Parmi les plus utiles, on retrouve notamment une fenêtre d'aide, de la 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 supportent un grande quantité de formats. Il

fournit également un support pour plusieurs projets ainsi qu'une interface pour utiliser des systèmes de contrôle des versions tels que GitHub (Horton, Kleinman, 2015).

4.8.3 Avantages et inconvénients de RStudio

- Exemples de fonctionnalités
  - Files, Plots, Packages, Help, and Viewer Pane Layout of the Components The RStudio interface consists of several main components sitting below a top-level toolbar and menu bar. Although this placement can be customized, the default layout utilizes four main panes in the following positions:

In the upper left is a Source browser pane for editing files (see Source Code Editor) or viewing some data sets. In Figure 1-3 this is not visible, as that session had no files open.

In the lower left is a Console for interacting with an R process (see Chapter 3).

In the upper right are tabs for a Workspace browser (see the section Workspace Browser) and a History browser (see the section Command History).

In the lower right are tabbed panes for interacting with the Files (The File Browser), Plots (Graphics in RStudio), Packages (Package Maintenance), and Help system components (The Help Page Viewer). If the facilities are present, an additional tab for version control (Version Control with RStudio) is presented.

The Console pane is somewhat privileged: it is always visible, and it has a title bar. For the other components, their tab serves as a title bar. These panes have page-specific toolbars (perhaps more than one)—which in the case of the Source pane are also context-specific.

The user may change the default dimensions for each of the panes, as follows. There is an adjustable divider appearing in the middle of the interface between the left and right sides that allows the user to adjust the horizontal allocation of space. Furthermore, each side then has another divider to adjust the vertical space between its two panes. As well, the title bar of each pane has icons to shade a component, maximize a component vertically, or share the space (Verzani, 2011). **Also check Nierhoff et Hillebrand 2015**

4.9 Comment utiliser RStudio ?

La première étape pour commencer à utiliser RStudio est de l'installer[2]. Une fois que cela est fait, ouvrez RStudio. La fenêtre qui apparait devrait ressembler à l'image ci-dessus. La couleur de l'arrière-plan et celle de la police, la taille des cadrans ainsi que de nombreux autres éléments peuvent être changés. La dernière section de ce chapitre aidera le lecteur à personnaliser son IDE.

Image de RStudio sans rien, settings de base.

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 quadran 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 et 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 a partir de RStudio.

[2] À cet effet, voir le chapitre 9 du présent ouvrage.

Image de RStudio qui montre les quatre cadrans. Idéalement avec un projet en cours et les différents cadrans utilisés. Settings personalisés?.

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 compose. 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 sans avoir à sortir de RStudio. L'onglet *Plots* permet de visualiser les graphiques générer à partir de R, que ce soit en utilisant *ggplot2, lattice ou base R*[3]. L'ongliet *Packages* permet de consulter les packages installés précédemment par l'utilisteur 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.

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

[3]Pour en apprendre davantage sur la visualisation graphique en R, consulter le chapitre 7 du présent ouvrage.

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 des l'espace de travail (Verzani, 2011; Nierhoff et Hillebrand, 2015).

4.10 Personnaliser son RStudio

One can easily switch between components using the mouse. As well, the View menu has subitems for this task. For power users, the keyboard shortcuts listed in Table 1-2 are useful. (A full list of keyboard shortcuts is available through the Help > Keyboard Shortcuts menu item.)

5 Baliser les sciences sociales : langages et pratiques

5.1 Question

Lorsque vous lisez une page Web, un article scientifique ou un curriculum vitæ professionnel, vous vous doutez peut-être que le texte n'est pas toujours produit à l'aide d'un simple logiciel de traitement de texte comme Microsoft Word, Apple Pages ou LibreOffice Writer. La mise en page réglée au millimètre près, la qualité des figures et graphiques, le style des références, la présence d'éléments interactifs et la cohérence hiérarchique du texte sont difficiles à reproduire à l'aide d'un logiciel de traitement de texte régulier, entre autres. L'insertion de tableaux de régression, de figures et d'extraits de code 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 et de la plateforme 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 tentera donc de répondre aux questions suivantes : « Pourquoi apprendre à utiliser des langages de balisage? Dans quels contextes sont-ils