Outils de recherche en sciences sociales numériques

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

2024-03-12

Table of contents

Avant-propos

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

Introduction

1 Données massives, causalité et sciences sociales : Changements et réflexions sur l'avenir

L'apparition des données massives (*big data*) dans le paysage technologique représente un cas de phénomène hautement technique dont les effets politiques et sociaux sont remarquables. Depuis quelques années, la discussion publique s'est en effet rapidement emparée du sujet, au point de transformer un développement 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?

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. Cela nous offre aussi l'opportunité d'aborder le sujet important de la différence entre les données expérimentales et observationnelles. Finalement, nous proposons quelques pistes de réflexion sur l'avenir des données massives en sciences sociales en identifiant certains changements *épistémologiques* que ces données pourraient potentiellement entraîner.

1.1 Définition des données massives

Il existe au moins trois approches conceptuelles permettant de définir les « données massives » (voir Figure 1.1.).

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. 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 entre données massives et données plus traditionnelles (ou « non-massives ») est simplement quantitative.

2. Deuxièmement, les données massives constituent un ***ensemble de pratiques*** de collecte, de traitement et d'analyse de ces points d'information. Les données massives représentent une technique, c'est-à-dire une manière ou une méthode nouvelle de faire de la recherche.

3. Finalement, d'une perspective sociologique, les données massives représentent les impacts sociaux de ces importants développements technologiques. Cette perspective souligne le caractère essentiellement social des données massives, en portant notamment attention aux risques liés à la confidentialité des données, aux enjeux relatifs au consentement et à l'autorisation de collecte des informations, aux innovations en intelligence artificielle, etc.

Dans les domaines scientifiques et technologiques, la définition courante attribuée 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 vélocité

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 (Kitchen et McArdle 2016). 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

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 relative à la différence entre les données expérimentales et les données observationnelles. Ces deux dimensions sont présentées de manière simultanées dans les prochaines sections.

1.2.1 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 en affirmant que des scores (y compris les résultats de classification qualitative) doivent capturer de manière significative les idées contenues dans le concept correspondant (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 le plus souvent représentatif de l'humeur de l'opinion publique. Ceci est tout aussi vrai pour les sondages à petits échantillons que pour ceux utilisant des données massives. 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, classifie les libertés civiles et les droits politiques des États du monde par degré afin de construire un index, ou une échelle, 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 la

question de la validité de la mesure est un élément central de ce qui unit (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 travaillant de manière indépendante les uns des autres).

1.2.2 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 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 l'utilisation des données massives permet désormais d'étudier. En se basant sur la collecte de plus de 11 millions de publications en ligne, King et ses collègues ont pu mesurer la censure exercée par le gouvernement chinois sur ces réseaux sociaux. En utilisant des données massives nouvelles, les auteurs 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; Bibeau et al. 2021).

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.

1.2.2.1 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*). Règle générale, 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, notamment parce qu'elle l'attribution aléatoire du traitement ou du contrôle est entièrement sous le contrôle du chercheur.e.s menant l'expérience. Cependant, en ce qui à trait à la plupart des données massives, elles sont générées de manière indépendante du contrôle du chercheur.e.s, et sont donc soumises aux mêmes enjeux et problèmes (biais) que les données observationnelles traditionnelles.

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.

1.2.2.2 Validité externe des données massives

Deuxièmement, les données massives représentent aussi un défi important pour la validité externe des recherches en sciences sociales (Tufekci 2014; Lazer et Radford 2017; Nagler et Tucker 2015). Un des problèmes les plus évidents concerne la ***représentativité*** des données massives collectées.

Comme le soulignent Lazer et Radford (2017), la quantité de données, en soi, 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. 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 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 enjeux 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.

Toutefois, depuis quelques années, le développement de nouvelles méthodes de pondération des données offre des pistes de solutions. La grande quantité de données massives permet notamment d'appliquer des méthodes de pondération bien plus efficaces pour corriger les échantillons non-représentatifs (Wang et al. 2015).

1.2.3 Données expérimentales

La question du processus de génération des données devient plus claire quand on considère comment les *données observationnelles* et les *données expérimentales* permettent d'effectuer des inférences de manière distincte (voir Figure 2). Toutefois, pour bien comprendre ce point, il faut comprendre les notions de données expérimentales et d'inférence causale, qui sont centrales au domaine de la causalité en recherche.

En quelques mots, l'essence de la démarche causale se résume comme suit : le processus de génération de données expérimentales a pour objectif d'assurer la validité d'une inférence causale estimée sur un échantillon sur l'ensemble de la population visée.

Plus spécifiquement, le processus de génération des données permet aux chercheur.e.s de s'assurer que la distribution du traitement entre les deux groupes, traitement et contrôle, est entièrement aléatoire. De manière technique, cette distribution aléatoire du traitement entre les deux groupes permet de garantir une source exogène (à l'opposé de endogène) de variation sur la variable indépendant (*x). Cette source exogène de variation*

*permet, quant à elle, d'éliminer l'endogénéité entre la variable indépendante (*x*) et le résidu (*e**).*

Autrement dit, le fait de distribuer au hasard le traitement entre les membres du groupe traitement et ceux du groupe contrôle assure que la variation dans les résultats ne vient pas d'autres facteurs non-contrôlés (le résidu, e), mais plutôt du traitement lui-même (la variable indépendante, x). En distribuant le traitement de manière aléatoire, on s'assure que les différences dans les résultats sont vraiment dues au traitement et non à d'autres facteurs.

Il s'agit là d'assurer le respect de la condition d'indépendance, essentielle à la validité de l'identification de l'effet causal étudié. Autrement dit, en éliminant l'endogénéité entre la x et e, on s'assure que l'effet observé n'est pas dû à une variable confondante.

Pour revenir aux données massives, celles-ci ne peuvent pas résoudre les enjeux liés aux inférences causales ou explicatives (Grimmer, 2015). Elles sont en effet également soumises aux mêmes impératifs issus du processus de génération des données.

1.2.4 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 récréer, de la manière la plus fidèle possible, une distribution aléatoire du traitement – ce que la littérature appelle un *as-if random assignment* (comme si l'attribution était aléatoire) (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.

Un autre exemple pourrait être l'utilisation du « matching », souvent utilisé dans les études économétriques. Supposons que vous souhaitez estimer l'effet d'un programme éducatif sur les résultats scolaires d'étudiant.e.s. Le devis de recherche idéal serait d'assigner aléatoirement les étudiant.e.s au programme (le groupe traitement) ou non (le groupe contrôle). Toutefois, puisque ce devis idéal peut être difficilement réalisable, un grand nombre de données pourrait permettre de trouver pour chaque étudiant.e dans le groupe traitement un étudiant.e « jumeau » dans le groupe contrôle. Ce « jumeau » serait similaire en âge, sexe, antécédents socio-économiques, etc. Il serait ensuite possible de comparer les résultats scolaires de ces jumeaux pour estimer l'effet du programme. Plus l'échantillon est grand, plus l'estimation sera précise et fiable, parce qu'il y aura plus de jumeaux possibles à apparier, réduisant ainsi le biais dû aux variables non observées.

Il s'agit d'un exemple où les données massives augmentent la validité interne de l'étude, même si les données sont de nature observationnelle et non expérimentale.

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 l'absence d'un biais de sélection — 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.

1.2.5 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 fidèlement 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.

	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 1.1: image2_2

1.3 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. L'analyse des données massives peut permettre de mettre en lumière des tendances subtiles échappant aux ensembles d'informations plus restreints.

Il faut toutefois souligner que les données massives représentent une complexification de l'analyse des phénomènes en sciences sociales d'une perspective non pas seulement méthodologique/technique mais également épistémologique.

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 mod-

erne?; (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 transparante: 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 pourrons 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 les avantages et les inconvénients, en plus de défis qui se posent. En guise de conclusion, nous souhaitons mettre l'accent 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

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 redistribuer[1]. 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 programme fait réellement. Seulement de cette façon, l'utilisateur peut *contrôler* le logiciel, plutôt que de se faire contrôler par ce dernier (Stallman 1986).

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

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 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.

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

À voir

Liste de logiciel libre répertorié par le gouvernement du Canada: https://code.open.canada.ca/fr/logiciels-libres.html#

Parallèlement au logiciel libre, il y a aussi le code ouvert, ou *open source*. A priori, la dénomination du logiciel libre et celle du *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 auxqels 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'ambiguité dans la dénomination "logiciel libre"[3], *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, 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:

[3]Soit ceux qui ont été conçus suivant les principes philisophiques et "moraux" qui sous-tendent ce mouvement.

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 license pour l'utilisation; 8) la license ne doit pas être spécifique pour un produit; 9) la license ne doit pas placer de restriction sur d'autres programmes; 10) la license doit être technologiquement neutre[4] (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 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 puiqu'ils ont des origines différentes, et qu'ils mènent à

[4]Pour plus d'informations sur ces caractéristiques, nous encourageons les lecteurs à se référer au lien web de Open Source Initiative (2006). Ils y trouveront un contenu détaillé pour chacunes des caractéristiques sus-mentionnées.

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

En quoi est-ce que ces deux concepts, issus du monde de l'informatique, sont-ils intéressants et/ou important pour les sciences sociales? Pour répondre à cette question, il est important de retourner à la base, soit de se questionner sur que sont les sciences sociales.

Ce chapitre à 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 souhaitions 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 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 [5], 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

[5]Allant de nos postulats ontologiques, épistémologiques et méthodologiques.

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 Inquiery*, 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. 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 2021, 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 réplicabilité ainsi que la qualité des recherches.

2.3.1 Avantages

2.3.1.1 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 crées une dynamique d'« innovation ascendante » et d'entraide (Couture 2014). Ce résultat constitue un avantage important 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 privé 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 raisonable 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 démocratise l'accès à plusieurs logiciels qui sont utiles pour mener des analyses scientifiques. Et ce, pour tous les utilisateurs dans le monde.

2.3.1.2 Avantages économiques : Une plus grande accessibilité pour tous

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 miliers de dollar. De plus, la license doit être renouvelée annuellement. Ce qui augmente les coûts associés à l'utilisation du logiciel et par conséquent limite son accessibilité. Comparativement, pour les logiciels libres, la license 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 Canada[6], les conditions d'emplois sont en ce moment[7] très bonnes, et une pénurie de main d'oeuvre est anticipé, entre 2022-2031, dans les emplois en analyse de données. 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 language 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. (Couture 2014; Karjalainen 2010).

[6]Ces informations proviennent du site web suivant: https://www.jobbank.gc.ca/marketreport/outlook-occupation/17882/ca

[7]En date d'écire ces lignes, septembre 2023.

2.3.2 Inconvénients et défis:

2.3.2.1 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 language de programmation afin de pouvoir utiliser le logiciel à son plein potentiel.

La synthaxe 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`.

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 synthaxe 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 language de programmation demande un investissement en temps, les bénéfices générées par ces nouvelles compétences dépassent le coût initial.

2.3.2.2 Problème de transparence

L'arrivé des sciences informatiques à fait émerger des problèmes de reproductibilité des protocoles scientifiques (Janssen 2017). Le problème principal 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 GitHub[8] 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 ignitié 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 transparent.

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.

[8] une plateforme publique *code ouvert* sur laquelle nous pouvons héberger et partager notre code.

Cependant, le partage du code utilisé reste encore majoritairement volontaire. Janssen, Pritchard, and Lee (2020) soutiennent que plus d'effort et d'actions concertés 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 2023). Par conséquent, rendre le code accessible devrait devenir un standard institutionnalisé.

2.3.2.3 Appropration capitaliste

Dans ce cas-ci, il s'agit plutôt d'un défis 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.

2.4 Critères de sélection

3 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.

3.1 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

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,

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 véridicité 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 disponible sur Windows, macOS et Linux.

3.2 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

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.

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

3.4 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 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*. L'onglet *Packages* permet de consulter les packages installés précédemment par l'utilisateur en plus de pouvoir en consulter la doc-

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

3.5 Conclusion

Le langage de programmation R est un outil très utile pour toutes sortes de tâches notamment reliées aux statistiques et à la visualisation graphiques. Sa maîtrise est requise pour accéder à plusieurs emplois, autant dans le monde académique que dans les secteurs publics et privés. Avec un peu de chance, le présent chapitre vous a éclairé sur son utilité et sa pertinence dans le monde du travail contemporain. Bien que le langage de programmation R ne doivent pas obligatoirement être utilisé avec RStudio, nous pensons que pour la plupart des usagers, leur utilisation conjointe est bénéfique et souhaitée. RStudio permet également d'utiliser différents langages de balisage compatibles avec R, facilitant l'utilisation de plusieurs outils complémentaires. L'apprentissage du langage de programmation R apparaît également être une valeur sure. Sa longitivité dans plusieurs sphères ainsi que la forte croissance de sa base d'usagers laisse présager que d'en connaître au moins les bases est un énorme avantage pour tout le monde. Pour ceux qui sont particulièrement intéressés par le langage de programmation R et qui désirent s'impliquer dans sa communauté, il

existe plusieurs conférences internationales et nationales sur R – notamment *RConference* and *useR!* – et un journal académique, *The R Journal*. On retrouve également différentes communautés telle que *R-Ladies* qui met de l'avant la diversité des genres dans la communauté du langage de programmation R. Le langage de programmation R est plus qu'un simple outil statistique, il est le centre d'une grande communauté de gens qui ont à coeur des principes liés à l'inclusion et à l'avancement humain.

4 À la quête de l'optimisation

Le monde de la recherche en sciences sociales numériques est en constante évolution, offrant de nouvelles opportunités mais aussi des défis uniques. Dans cette quête incessante pour optimiser notre efficacité et notre collaboration, l'utilisation des bons outils devient la clé de la réussite. Que vous soyez un chercheur en herbe ou un professionnel chevronné, la manière dont vous organisez vos méthodes de travail et gérez vos ressources peut déterminer la qualité et l'impact de vos résultats.

4.1 L'importance d'une méthode de travail efficace

Avant même de plonger dans les détails des méthodes de recherche et des analyses, il est crucial de poser les bases d'une méthode de travail efficace. Qu'il s'agisse de travailler en solitaire ou en équipe, l'ordre et la structure sont des éléments essentiels. Des dossiers bien organisés, une arborescence claire et un entreposage sécurisé deviennent les piliers sur lesquels repose votre productivité. Après tout, un environnement de travail organisé engendre des résultats ordonnés.

Ce chapitre vous emmènera à découvrir une gamme d'outils conçus pour répondre aux besoins spécifiques des chercheurs en sciences sociales numériques. Dans une quête pour maximiser votre temps, améliorer vos flux de travail et renforcer vos collaborations, nous explorerons trois types d'outils qui vous guideront dans cette quête d'optimisation :

1. **Logiciels de communication** : La communication transparente est le cœur d'une collaboration réussie. Nous explorerons des outils tels que Slack qui facilitent les échanges en temps réel, connectant les chercheurs, même à distance, pour un partage rapide d'idées et d'informations.

2. **Logiciels de gestion de versions décentralisé** : Nous plongerons dans le monde de Git et GitHub, des outils indispensables pour le suivi des versions et la collaboration efficace sur le code source.

3. **Outils d'entreposage de données** : Que vous traitiez des données sensibles ou non, la conservation sécurisée de vos informations est primordiale. Des plateformes telles que Dropbox et Amazon Web Services (AWS) offrent des espaces sécurisés pour entreposer et partager vos données avec votre équipe.

Chacun de ces outils est une pièce du puzzle, conçue pour vous aider à gagner du temps, à collaborer de manière plus fluide et à renforcer la qualité de votre recherche en sciences sociales numériques. Plongeons dans ces outils avec un désir commun d'optimisation et d'excellence dans notre travail.

4.2 Logiciel de gestion de communication (Slack)

Dans tout bon projet de recherche, la communication est primordiale. Que ce soit pour décrire les avancements, discuter des étapes à venir, entretenir un partenariat avec des partenaires ou simplement structurer ses pensées, la plateforme par laquelle vous communiquez vous accompagne à chacune des étapes du travail. Il est donc important de choisir un outil qui convient bien à vos projets et de prendre le temps de l'apprivoiser et d'optimiser son utilisation.

Il y a tellement de plateformes différentes pour communiquer qu'il faut être prudent par rapport au nombre utilisé. Si vous ne faites pas un choix,

vous pouvez, sans vous en rendre compte, mêler Teams, courriels, Zoom et autres. Rapidement, vous perdez le contrôle de ce qui est dit. Nous vous proposons d'opter pour un logiciel de gestion de communication. Il existe plusieurs logiciels du genre, tels que Microsoft Teams, Slack, Google Workspace et Workplace. Toutes ces options peuvent vous permettre de collaborer efficacement en équipe. Dans le cadre de nos travaux, nous utilisons Slack. C'est donc principalement de cet outil que nous parlerons dans cette section, mais n'hésitez pas à vérifier quelle plateforme correspond le mieux à vos besoins.

4.2.1 Pourquoi utiliser une de ces plateformes

Peu importe votre niveau d'implication, la collaboration et la communication sont inévitables en recherche. La science n'est pas une discipline qui se développe en solitaire, elle nécessite des échanges et des débats. Les équipes de recherche sont souvent dispersées géographiquement. Même si vous travaillez actuellement seulement avec votre directeur, il est certain que plusieurs équipes de recherche dans votre département utilisent un tel logiciel. Un courriel peut faire l'affaire pour une discussion ponctuelle qui se règle rapidement. Cependant, dans une équipe de travail dynamique, où plusieurs membres participent à divers projets, les courriels deviennent rapidement chaotiques, il est difficile de retracer ce qui a été dit et de conserver les pièces jointes. Les discussions deviennent rapidement trop complexes pour le médium utilisé.

Les logiciels de gestion de communication ont été conçus spécifiquement pour répondre aux besoins des équipes collaboratives. Vous y trouverez leur facette la plus attrayante : une structure simple et adaptée. Les chaînes et fils de discussions permettent de garder des traces et de se retrouver facilement dans ce qui a été dit. Une autre force de ces logiciels est la centralisation des outils de travail. Sur Slack, comme sur Teams, vous pouvez faire des appels en visioconférence à l'endroit où vos conversations

écrites se trouvent. Il est aussi possible d'y télécharger l'application mobile, ce qui facilite l'accessibilité et la connexion des membres de l'équipe. Tout avoir structuré à son goût au même endroit et à portée de main, cela permet de structurer sa pensée plus efficacement, d'éviter les oublis et de réduire le stress.

4.2.2 Comment utiliser votre logiciel efficacement

Une fois que vous êtes convaincu d'aller de l'avant avec un de ces outils, vous devrez apprendre à bien vous en servir. Voici quelques trucs qui pourront vous aider à optimiser son utilisation. Les points ci-dessous font référence à Slack, mais peuvent très bien être adaptés à d'autres plateformes.

4.2.2.1 Structuration

Il est important de bien réfléchir à la structuration de vos chaînes. Si vous ne faites pas ce travail, les chaînes peuvent se multiplier rapidement et les conversations se mettent alors à s'entrecroiser, vous faisant ainsi perdre le fil. L'objectif de ces outils étant d'évitez ces problèmes, vous ne voulez pas perdre l'avantage comparatif que vous venez tout juste de gagner face aux courriels! La structuration des chaînes devrait être similaire à celle de votre équipe de recherche. Si vous utilisez Notion ou un autre logiciel du genre, la structure des deux outils devrait être la même. Nous vous proposons d'avoir une chaîne pour chacun des projets. Si le projet est trop gros et que la conversation devient chaotique, pensez à créer une sous-chaîne (un sous-projet) qui vous permettra d'aborder un sujet précis, sans mêler les discussions. Pour faciliter la structuration des chaînes, vous pouvez utiliser des préfixes, pour classer les chaînes par thème, ou autre typologie qui vous convient. Également, utilisez les Espaces d'équipe. Chaque équipe devrait avoir son propre espace, avec ses propres chaînes. Vous pouvez faire partie de plusieurs équipes et naviguer à travers les espaces. Si plusieurs équipes

partagent un même espace de travail, vous pourriez perdre le contrôle de sa structure.

4.2.2.2 Maintenance

Slack est un espace dynamique, tout comme votre équipe! La structure que vous avez choisie n'est pas permanente. Vous devriez rapidement vous questionner à savoir si elle convient toujours à vos activités. Votre espace d'équipe est comme votre réel lieu de travail, faites-y régulièrement le ménage pour vous assurer que tout est propre et en ordre. Archivez les chaînes qui ne sont plus pertinentes ou actives, puisque vous pourrez toujours les désarchiver quand cela sera nécessaire. Épinglez des messages importants et des documents utiles aux projets dans les chaînes appropriées. Faites le tour de ce qui est épinglé à l'occasion pour vérifier si c'est encore pertinent. Cela peut paraître énergivore, mais l'efficacité de votre travail d'équipe va en bénéficier. Également, rappelez aux membres de votre équipe d'utiliser les bonnes chaînes pour chacune des discussions. Il ne faut pas que les conversations se croisent à travers les chaînes. Chaque chaîne a son utilité et doit être utilisée en conséquence. Les appels d'équipe doivent aussi se faire dans les bonnes chaînes. Quand vous êtes en appel, utilisez le fil de discussion pour conserver des traces écrites des points abordés dans la réunion. Les fils de discussions sont en général un bon outil pour ne pas se perdre dans un discussion. Si l'usage des mauvaises chaînes est un problème récurent, il est possible que la structure que vous employez est mal adaptée à vos travaux. Vous pouvez alors retourner à la planche à dessin. Assurez-vous que toute l'équipe comprenne bien comment utiliser Slack. Si ce n'est pas le cas, formez-les. Une structure adaptée et une équipe bien formée peuvent faire des miracles.

4.2.2.3 Collaboration

La grande majorité des conversations devraient se faire dans les chaînes. Les conversations privées ont leur utilité, vous vous en servirez. Il est parfois nécessaire d'avoir des discussions plus confidentielles et de parler rapidement à quelqu'un sur un sujet éphémère. Toutefois, par soucis de transparence et d'inclusion, toute discussion à propos d'un projet devrait se faire dans sa chaîne. Si vous jugez qu'un membre d'une chaîne ne devrait pas lire ce que vous avez à dire sur le projet, c'est qu'il ne devrait pas faire partie de la chaîne. Par rapport aux membres, trouvez le bon équilibre par rapport à qui devrait être dans quelle chaîne. L'objectif n'est pas d'exclure et de cacher du contenu, vous voulez une équipe transparente. Vous voulez que vos membres restent bien informés de l'avancement des projets sans les submerger d'information qui ne leur est pas utile. C'est à vous de trouver la formule gagnante. Invitez vos partenaires externes dans votre espace d'équipe. Créez des chaînes spécifiques aux partenaires pour que les conversations externes soient tout aussi organisées. N'invitez pas vos partenaires dans vos chaînes privées, question de confidentialité. Si un partenaire n'a pas l'habitude d'utiliser Slack ou l'outil que vous utilisez, proposez-lui de vous y joindre quand même. Moins vous utilisez les outils des autres, plus vous gardez centralisées vos communications et évitez de jongler avec plusieurs plateformes.

4.2.2.4 Optimisation personnelle

Une fois que la structure d'équipe est définie et que vos membres et vos partenaires sont à l'aise avec l'utilisation de la plateforme, il est temps d'organiser la structure de votre Slack personnel. Créez des sections pour trier les chaînes. La structure d'équipe est essentielle, mais une fois qu'elle est déterminée, chaque membre n'utilise pas forcément les chaînes de la même façon. Vous pouvez vous créer une section de favoris, ou encore différentes sections par rapport aux différents thèmes pour y faciliter la navigation. Également, ajustez vos paramètres de notifications. C'est à

vous de déterminer quelle chaînes méritent de produire des alertes, et à quels moments vous souhaitez les recevoir. Slack a plusieurs applications intégrées qui facilitent la compatibilité avec vos autres outils. Vous pouvez connecter votre calendrier, votre Notion et votre GitHub pour recevoir des alertes pertinentes. Allez explorer ces applications pour déterminer lesquelles vous conviennent.

Tel que mentionné précédemment, plusieurs logiciels peuvent convenir à vos besoins. Puisque nous utilisons Slack, voici quelques raisons qui pourraient vous convaincre d'opter pour cette option ou de vous en éloigner. Sachez que cette liste n'est pas du tout exhaustive, mais reflète simplement quelques-unes de nos observations par rapport à notre outil de travail.

- Avantages

  L'utilisation de Slack est très intuitive. Nous l'utilisons régulièrement dans des cours, et les étudiants apprennent rapidement à l'utiliser. La distinction entre les chaînes publiques accessibles à tous les membres d'un espace d'équipe et les chaînes privées est claire et simple d'utilisation. Slack offre aussi une fonction de recherche, qui vous permet de retrouver des messages à travers les chaînes. L'intégration de applications qui font le pont avec d'autres outils est fort appréciée. Enfin, Slack est utilisé partout dans le monde par des équipes de toutes les tailles et dans tous les domaines. C'est un outil très présent en recherche académique qui facilite la collaboration et la multidisciplinarité. Les chances sont élevées que vos partenaires utilisent déjà l'outil, ou au minimum en aient déjà entendu parlé.

- Inconvénients

  Si vous avez l'habitude d'utiliser les outils d'une suite, comme celles de Microsoft ou de Google, il est possible que vous trouviez l'intégration de ces outils à Slack moins pratique que si vous utilisiez les plateformes proposées par ces compagnies. Également, gardez en tête que la version gratuite de Slack a plusieurs limitations. Elle

implique notamment un limite de temps par rapport à l'archivage des messages, que vous ne pourrez pas retracer après 90 jours. Les coûts pour utiliser Slack à son plein potentiel peuvent être élevés, mais puisque ce genre d'outils est de plus en plus répandu, il est fort possible que son utilisation soit financée par votre département.

4.3 Logiciel de gestion de versions décentralisé

Lorsque l'on aborde le domaine de la recherche scientifique en sciences sociales numériques, la collaboration et la gestion efficace du code deviennent des éléments cruciaux pour progresser dans ses projets. Dans cette optique, les outils de gestion de versions décentralisés ont pris une place prépondérante. Parmi eux, Git et GitHub se démarquent tant par leur popularité que par leur efficacité.

4.3.1 Pourquoi choisir Git et GitHub?

4.3.1.1 Avantages

Git, développé par Linus Torvalds en 2005, s'est imposé comme le système de gestion de versions décentralisé de référence. Sa principale force réside dans sa capacité à suivre l'évolution d'un projet en enregistrant les modifications apportées au code source. Chaque modification est enregistrée sous forme de dépôts (*commits*), avec un message explicatif, permettant aux collaborateurs de comprendre facilement les évolutions du projet.

GitHub, lancé en 2008, est une plateforme qui utilise Git comme base pour l'entreposage et la gestion de projets. C'est une vitrine virtuelle où les développeurs peuvent héberger leurs dépôts Git et collaborer de manière transparente. L'aspect social de GitHub, avec ses fonctionnalités de suivi des projets, de gestion des problèmes et de demandes de fusion, en fait un lieu de choix pour les projets en code source ouvert et collaboratifs.

En sciences sociales numériques, où le partage et la collaboration sont essentiels, Git et GitHub offrent plusieurs avantages majeurs. Tout d'abord, ils permettent de suivre les modifications apportées au code, ce qui facilite la reproductibilité des résultats. Les chercheurs peuvent revenir à n'importe quelle version précédente du code, ce qui est particulièrement utile pour corriger des erreurs ou analyser l'impact de différentes approches.

De plus, Git et GitHub favorisent le travail collaboratif. Plusieurs chercheurs peuvent travailler sur le même projet simultanément, chacun dans sa branche de développement. Une fois les modifications effectuées, il est possible de fusionner les branches pour intégrer les changements. Cette approche évite les conflits majeurs et facilite la répartition des tâches au sein de l'équipe.

Enfin, l'aspect de code source ouvert de GitHub permet aux chercheurs en sciences sociales numériques de partager leurs codes avec la communauté académique et de bénéficier des contributions d'autres chercheurs. Cela favorise un environnement de partage des connaissances et de collaboration fructueuse.

4.3.1.2 Inconvénients

Cependant, Git et GitHub ne sont pas sans leurs défis. La courbe d'apprentissage peut être raide pour les débutants, car ces outils impliquent des concepts spécifiques tels que les branches, les conflits de fusion et les requêtes de tirage. De plus, bien que GitHub offre un niveau de gratuité pour les projets en code source ouvert, des frais peuvent être appliqués pour des fonctionnalités avancées ou pour des projets privés.

4.3.2 Comment les utiliser efficacement (en parallèle à Dropbox, etc.)

Pour utiliser Git et GitHub efficacement dans un contexte de recherche en sciences sociales numériques, il est recommandé de suivre quelques bonnes pratiques. Tout d'abord, il est important de structurer son dépôt Git de manière logique, en organisant les fichiers et les dossiers de manière cohérente. Les messages de commit doivent être descriptifs et clairs, pour permettre à tous les collaborateurs de comprendre les changements effectués.

Il est également conseillé de travailler sur des branches distinctes pour chaque fonctionnalité ou modification majeure. Cela facilite la gestion des changements et minimise les conflits lors de la fusion. Les chercheurs devraient également consulter régulièrement les projets et les problèmes sur GitHub pour encourager une communication ouverte et résoudre rapidement les problèmes.

L'utilisation de Git et de GitHub peut être complémentaire à d'autres outils d'entreposage, tels que Dropbox ou Google Drive. Ces derniers peuvent être utilisés pour entreposer des fichiers non liés au code, tels que des données brutes non sensibles ou des documents de recherche, tandis que Git et GitHub gèrent le code source et ses évolutions.

Bien qu'il existe plusieurs alternatives à l'utilisation combinée de Git et de GitHub sur le marché, ces deux plateformes liées continuent de dominer le domaine de la gestion de versions décentralisée. Parmi les alternatives notables, on peut citer Mercurial, Bitbucket, GitLab et SourceForge. Chacun de ces outils offre des fonctionnalités similaires à celles de Git et GitHub, mais il est important de comprendre pourquoi Git et GitHub restent les choix privilégiés pour les chercheurs en sciences sociales numériques.

4.3.3 Pourquoi prioriser Git et GitHub pour les chercheurs en sciences sociales

1. *Intégration et adoption répandue* : Git est devenu un standard de facto dans l'industrie du développement logiciel. Sa popularité et son adoption répandue signifient que de nombreuses ressources d'apprentissage, des tutoriels et des forums de support sont disponibles en ligne, ce qui facilite l'utilisation de cet outil pour les chercheurs en sciences sociales débutants. GitHub, en tant que plateforme principale de gestion des versions, bénéficie également d'une grande base d'utilisateurs et d'une communauté active, ce qui encourage la collaboration et le partage des connaissances.

2. *Facilité de collaboration* : Git et GitHub sont conçus pour faciliter la collaboration entre les individus et les équipes. Les chercheurs en sciences sociales travaillent souvent ensemble sur des projets de recherche, et la capacité de suivre les modifications, de gérer les conflits et de fusionner les contributions devient essentielle. L'interface conviviale de GitHub, avec des fonctionnalités telles que les demandes de fusion et les commentaires en ligne, simplifie grandement la collaboration.

3. *Visibilité et partage* : GitHub brille par sa fonctionnalité de projet open source, qui permet aux chercheurs en sciences sociales de partager leurs travaux avec la communauté mondiale. Les projets en code source ouvert sont visibles et accessibles à tous, favorisant ainsi la collaboration et l'examen par les pairs. Cela peut être particulièrement bénéfique pour les chercheurs souhaitant contribuer à des initiatives académiques et collaborer à des projets interdisciplinaires.

4. *Suivi des versions et recherche reproductible* : Les chercheurs en sciences sociales doivent s'assurer que leurs travaux sont reproductibles et vérifiables. Git permet de suivre les versions du code, ce qui signifie que les chercheurs peuvent retrouver facilement des versions antérieures pour reproduire des analyses spécifiques ou corriger des

erreurs. Cette fonctionnalité est cruciale pour maintenir l'intégrité des résultats de recherche.

5. *Infrastructure et sécurité* : GitHub offre une infrastructure robuste pour l'entreposage sécurisé des dépôts Git. Les chercheurs peuvent être assurés que leurs travaux sont sauvegardés et protégés contre les pertes de données accidentelles. De plus, les contrôles d'accès et les autorisations granulaires de GitHub permettent aux chercheurs de contrôler qui peut accéder et contribuer à leurs projets.

En somme, Git et GitHub offrent aux chercheurs en sciences sociales numériques un moyen puissant de gérer leur code, de collaborer efficacement et de contribuer à la communauté académique grâce à l'open source. Bien que leur apprentissage puisse représenter un défi initial, les avantages qu'ils apportent en termes de suivi des versions, de collaboration et de partage des connaissances en font des outils essentiels dans l'arsenal de tout chercheur moderne.

4.3.4 Pratiques à éviter sur GitHub pour les chercheurs en sciences sociales

Lorsque les chercheurs en sciences sociales utilisent GitHub pour partager leur code, collaborer sur des projets et contribuer à la communauté académique, il est essentiel de connaître les pratiques à éviter. En effet, certaines erreurs peuvent compromettre la sécurité, la confidentialité et l'efficacité de la recherche. Voici quelques éléments à éviter :

1. *Entreposer des informations sensibles* : Évitez d'entreposer des données sensibles ou confidentielles sur GitHub. Cela inclut les données de sondages, les informations personnelles identifiables et tout autre contenu pouvant porter atteinte à la vie privée des individus. Assurez-vous de supprimer ou de masquer soigneusement ces informations avant de les télécharger sur la plateforme.

2. *Inclure des mots de passe et clés d'accès* : Ne jamais inclure de mots de passe, de clés d'accès ou d'informations d'identification dans votre code source. Cela peut compromettre la sécurité de vos systèmes et de vos données. Utilisez plutôt des méthodes sécurisées pour gérer ces informations, telles que les variables d'environnement ou les fichiers de configuration externes.

3. *Entreposer des fichiers lourds* : Évitez d'entreposer des fichiers volumineux sur GitHub, notamment des fichiers binaires, des données brutes massives ou des ensembles de données volumineux. Ces fichiers peuvent ralentir les opérations de clonage et de fusion, ce qui affecte la performance globale du dépôt. Utilisez plutôt des services d'entreposage dédiés pour ces fichiers et fournissez des liens vers ces ressources dans votre dépôt.

4. *Inclure des identifiants personnels* : Évitez de publier vos propres identifiants personnels, tels que des numéros de sécurité sociale, des numéros de carte de crédit ou d'autres informations confidentielles. Ces informations pourraient être exploitées à des fins malveillantes si elles tombent entre de mauvaises mains.

5. *Ignorer les pratiques de branches et de fusion* : Évitez de fusionner directement du code dans la branche principale (habituellement appelée *main* ou *master*). Utilisez plutôt des branches distinctes pour les fonctionnalités et les corrections, et suivez les pratiques de fusion pour intégrer proprement les changements. Ignorer ces pratiques peut entraîner des conflits et une perte de trace des modifications.

6. *Ignorer les commentaires des collaborateurs* : Lorsque vous travaillez avec d'autres chercheurs, ne négligez pas les commentaires et les suggestions qu'ils fournissent. Les retours d'expérience et les idées des autres peuvent contribuer à améliorer la qualité de votre code et de vos analyses.

7. *Ne pas documenter* : Évitez de ne pas documenter votre code. Une documentation claire et détaillée est essentielle pour permettre à

d'autres chercheurs de comprendre vos méthodes et vos résultats. Utilisez des commentaires explicatifs et fournissez des explications sur la manière d'exécuter votre code.

En suivant ces conseils et en évitant ces erreurs courantes, les chercheurs en sciences sociales peuvent garantir la sécurité, la qualité et l'efficacité de leurs projets sur GitHub. La responsabilité de préserver la confidentialité des données et de créer un environnement de travail collaboratif et respectueux repose sur les épaules de chaque contributeur.

4.3.5 Exemple d'utilisation de Git et de GitHub pour un chercheur en sciences sociales

Dans le contexte de la recherche en sciences sociales numériques, la gestion efficace du code, la collaboration transparente et la préservation des données sensibles sont des impératifs. Imaginons que vous êtes un jeune chercheur en sciences sociales qui étudie l'impact des médias sur l'opinion publique. Vous utilisez le langage de programmation R pour analyser des données de médias et des données de sondage. Bien que vous travailliez seul, vous souhaitez rendre votre travail accessible à votre équipe pour validation et permettre à vos collègues de contribuer aux améliorations. Voici comment vous pouvez utiliser Git et GitHub pour gérer votre projet de manière structurée et collaborative.

4.3.5.1 Étape 1 : Création d'un répertoire local et initialisation de Git

Ouvrez votre terminal et naviguez vers le dossier où vous souhaitez enregistrer votre projet.

```
cd chemin/vers/votre/dossier
```

Créez un nouveau répertoire pour votre projet et accédez-y.

```
mkdir mon_projet
```

```
cd mon_projet
```

Initialisez Git dans ce répertoire.

```
git init
```

4.3.5.2 Étape 2 : Ajout de votre code et de vos fichiers

Ajoutez vos fichiers R contenant le code pour l'analyse des médias et des sondages dans le répertoire. Par exemple, vous pouvez avoir des fichiers *analyse_medias.R* et *analyse_sondages.R*.

Utilisez la commande `git status` pour vérifier l'état de vos fichiers.

```
git status
```

4.3.5.3 Étape 3 : Ajout, validation et commit de vos modifications

Ajoutez vos fichiers pour qu'ils soient prêts à être validés.

```
git add -A
```

Validez vos modifications avec un message descriptif.

```
git commit -m "Ajout du code d'analyse des médias et des sondages"
```

4.3.5.4 Étape 4 : Création du répertoire sur GitHub et du lien avec votre répertoire local

Allez sur GitHub et connectez-vous à votre compte. Créez un nouveau répertoire vide avec le nom *mon_projet*.

De retour dans votre terminal, ajoutez le lien GitHub à votre répertoire local.

```
git remote add origin https://github.com/votre-utilisateur/mon_projet.git
```

4.3.5.5 Étape 5 : Push de votre travail sur GitHub

Envoyez vos commits locaux vers GitHub.

```
git push -u origin master
```

4.3.5.6 Étape 6 : Collaboration avec vos collègues

Si vos collègues souhaitent contribuer à votre projet, ils peuvent *forker* votre répertoire sur GitHub, ce qui créera une copie dans leur propre compte.

Lorsqu'ils ont fait des modifications dans leur copie, ils peuvent soumettre une *pull request* pour vous demander de fusionner leurs modifications dans votre répertoire principal.

4.3.5.7 Étape 7 : Pull des modifications de vos collègues

Lorsque vos collègues ont soumis des modifications et vous ont demandé de les fusionner, vous pouvez mettre à jour votre répertoire local avec leurs changements.

```
git pull origin master
```

4.3.5.8 Étape 8 : Répéter le processus

Répétez les étapes 2 à 7 au fur et à mesure que vous développez votre projet, ajoutez du code, effectuez des analyses et collaborez avec vos collègues. Assurez-vous de valider et de pousser régulièrement vos modifications pour maintenir le dépôt à jour.

4.3.6 GitHub Desktop

Alors que le terminal reste une approche fondamentale pour maîtriser Git et GitHub, il existe des outils conviviaux tels que GitHub Desktop qui offrent une alternative intuitive. Cet outil simplifie le processus de gestion de versions décentralisée, en particulier pour ceux qui souhaitent commencer par une approche visuelle. Cependant, comprendre son fonctionnement et équilibrer les avantages et les inconvénients est essentiel.

GitHub Desktop fournit une vue claire de vos dépôts, de vos modifications, de vos branches et de vos demandes de fusion. Il élimine la nécessité de mémoriser les commandes en ligne de terminal, ce qui peut être un défi pour certains chercheurs. L'application simplifie également la résolution des conflits lors de la fusion des branches.

Toutefois, en utilisant GitHub Desktop, il est possible de perdre la compréhension des commandes Git en ligne de commande, ce qui pourrait devenir un inconvénient si vous devez travailler dans un environnement sans interface visuelle. De plus, GitHub Desktop est spécifiquement conçu pour interagir avec GitHub. Si vous devez travailler avec d'autres plateformes de gestion de versions, cela pourrait poser des problèmes.

La décision entre l'utilisation du terminal et de GitHub Desktop dépend de vos préférences et de vos besoins. Pour les chercheurs qui débutent,

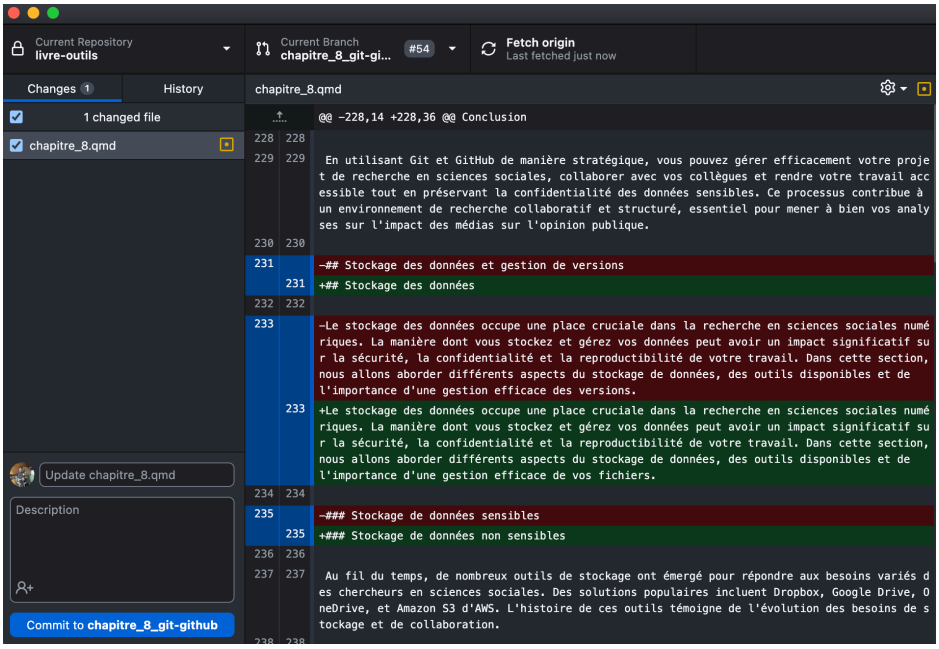


Figure 4.1: image

GitHub Desktop offre une transition en douceur vers les concepts de gestion de versions. Cependant, il est important de ne pas se limiter à une interface visuelle. Comprendre les commandes Git en ligne de commande reste essentiel pour résoudre des problèmes complexes, gérer des projets avancés et collaborer avec d'autres chercheurs qui utilisent des approches basées sur le terminal.

4.4 Conclusion

En utilisant Git et GitHub de manière stratégique, vous pouvez gérer efficacement votre projet de recherche en sciences sociales, collaborer avec vos collègues et rendre votre travail accessible tout en préservant la confidentialité des données sensibles. Ce processus contribue à un environnement de recherche collaboratif et structuré, essentiel pour mener à bien vos analyses sur l'impact des médias sur l'opinion publique.

4.5 Outils d'entreposage des données

L'entreposage des données occupe une place cruciale dans la recherche en sciences sociales numériques. La manière dont vous entreposez et gérez vos données peut avoir un impact significatif sur la sécurité, la confidentialité et la reproductibilité de votre travail. Dans cette section, nous allons aborder différents aspects de l'entreposage de données, des outils disponibles et de l'importance d'une gestion efficace de vos fichiers.

4.5.1 Entreposage de données non sensibles

Au fil du temps, de nombreux outils d'entreposage ont émergé pour répondre aux besoins variés des chercheurs en sciences sociales. Des solutions populaires incluent Dropbox, Google Drive, OneDrive et Amazon

S3 d'AWS. L'histoire de ces outils témoigne de l'évolution des besoins d'entreposage et de collaboration.

Lorsqu'il s'agit d'entreposer vos données de recherche, la règle d'or est de ne jamais perdre d'informations précieuses. Cette préoccupation prend toute son importance lorsqu'un chercheur en sciences sociales, seul ou en équipe restreinte, se lance dans un projet. Pour répondre à ce besoin, les services d'entreposage cloud tels que Dropbox, Google Drive et OneDrive se révèlent indispensables. Voici quelques avantages d'un entreposage sur le cloud pour la recherche :

1. *Sauvegarde automatique* : Les solutions cloud sauvegardent automatiquement vos fichiers, garantissant que vous ne perdrez jamais vos données en cas de panne d'ordinateur ou d'accident.

2. *Accessibilité universelle* : Vous pouvez accéder à vos fichiers à partir de n'importe quel appareil avec une connexion Internet, ce qui favorise la flexibilité dans la gestion de vos projets.

3. *Partage facilité* : Les services cloud permettent de partager facilement des fichiers et des dossiers avec des collègues, même en dehors de votre équipe de recherche. Cela favorise la collaboration et la communication.

Il est important de noter que le choix d'un service cloud dépend de vos besoins et de vos préférences. Considérez des facteurs tels que la capacité d'entreposage, les fonctionnalités de partage, la convivialité et la compatibilité avec vos outils de recherche existants.

Dropbox est connu pour sa simplicité d'utilisation et sa convivialité. Il peut être un choix approprié pour entreposer des fichiers non sensibles, partager des documents avec des collègues et faciliter la collaboration.

Pour utiliser Dropbox efficacement, organisez vos fichiers en arborescence logique. Créez des dossiers spécifiques pour chaque projet et partagez-les avec les membres de votre équipe. Pour éviter de pousser des fichiers

sensibles sur GitHub, ajoutez le nom de dossier à exclure dans un fichier *.gitignore*.

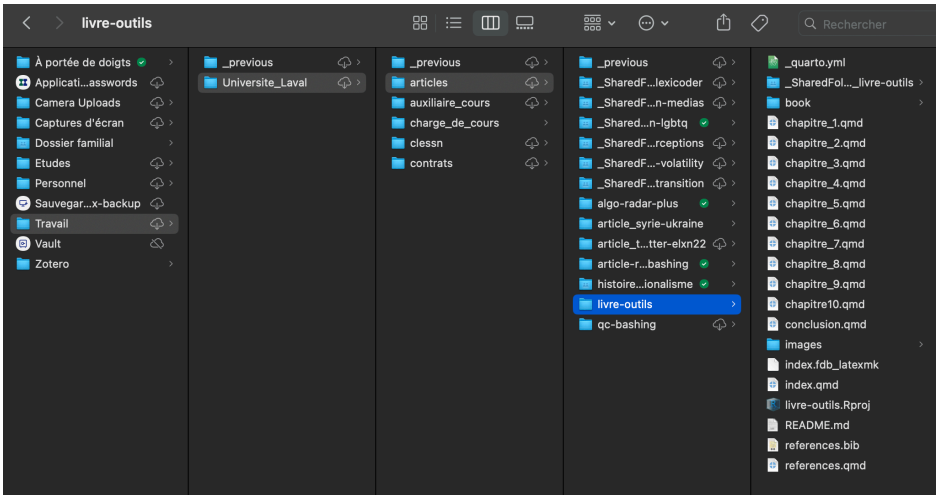


Figure 4.2: image1

Dropbox offre un suivi automatique des modifications, ce qui vous permet de remonter dans le temps pour restaurer des versions antérieures de vos fichiers. Cela garantit l'intégrité de vos données et vous permet de revenir à des versions précédentes si nécessaire. De plus, l'archivage de dossiers et de projets complets peut aider à conserver une vue chronologique de votre travail au fil du temps.

Il est également crucial de considérer la taille de vos données. Si vous traitez des fichiers volumineux tels que des images, des vidéos ou des ensembles de données massifs, il peut être judicieux d'utiliser un service cloud pour entreposer ces fichiers et les partager avec vos collaborateurs, plutôt que de les pousser sur des plateformes de gestion de versions comme GitHub.

Pour les données sensibles, les services cloud tels que Dropbox et Google Drive peuvent ne pas être suffisamment sécurisés. C'est là que des solu-

tions comme AWS entrent en jeu. Cependant, il est important de noter que l'utilisation d'AWS peut s'avérer complexe, en particulier pour un jeune chercheur travaillant en solo ou en petite équipe.

4.5.2 Entreposage de données sensibles

Lorsqu'il s'agit d'entreposer des données sensibles, tels que des données de sondage comportant des informations personnelles identifiables, la sécurité et la confidentialité sont essentielles. Comme abordé précédemment, GitHub n'est pas adapté à l'entreposage de telles données en raison de ses caractéristiques publiques et de son orientation vers le code source ouvert. Une solution courante est d'utiliser des services de cloud sécurisés, tels qu'AWS, qui offrent des mesures de sécurité robustes pour protéger vos données sensibles.

AWS regroupe un ensemble de services *cloud* proposés par Amazon. Il offre une vaste gamme de services, allant de l'entreposage et de la gestion des données à la computation et à l'analyse avancée. AWS est conçu pour offrir une infrastructure hautement évolutive et sécurisée, ce qui en fait un choix attrayant pour les chercheurs qui gèrent des données sensibles. L'outil présente de multiples avantages:

1. *Sécurité robuste* : AWS met l'accent sur la sécurité, avec des fonctionnalités telles que le chiffrement des données en transit et au repos, la gestion des accès basée sur les rôles et la conformité à des normes de sécurité strictes.

2. *Scalabilité* : AWS permet de faire évoluer vos ressources en fonction des besoins, garantissant des performances optimales même lorsque vos projets de recherche croissent en taille et en complexité.

3. *Flexibilité* : AWS propose une variété de services adaptés à différentes utilisations, allant de l'entreposage de données au calcul intensif pour l'analyse avancée.

4. *Collaboration simplifiée* : Bien que le coût d'entrée soit généralement bas, la possibilité de partager des ressources avec des collègues et de travailler en équipe rend AWS adapté à la collaboration.

AWS n'est pas le seul service cloud disponible. Microsoft Azure et Google Cloud Platform (GCP) sont des concurrents majeurs offrant des fonctionnalités similaires. Lorsque vous choisissez un fournisseur, prenez en compte les coûts, la convivialité et les fonctionnalités offertes. Le coût d'utilisation d'AWS peut varier en fonction des services utilisés, de la quantité de données entreposées et de la capacité de calcul requise. Lorsque vous travaillez seul, le coût peut sembler élevé par rapport à l'utilisation de solutions gratuites telles que Dropbox. Cependant, en équipe, la répartition des coûts peut rendre AWS plus abordable.

4.5.2.1 Exemple d'utilisation d'AWS pour entreposer et accéder à des données de sondages dans RStudio

Imaginez un jeune chercheur en sciences sociales qui travaille sur une analyse comparative de données de sondages recueillies sur plusieurs décennies. Pour maintenir la sécurité des données sensibles et faciliter l'accès pour les analyses dans RStudio, il décide d'utiliser AWS pour l'entreposage et la gestion de ses données.

4.5.2.1.1 Étape 1 : Création d'un compte AWS et configuration

Le chercheur crée un compte AWS et configure ses paramètres de sécurité, y compris la configuration de l'authentification à deux facteurs pour renforcer la sécurité de son compte.

4.5.2.1.2 Étape 2 : Création d'un espace d'entreposage S3

Le chercheur crée un compartiment Amazon S3 (Simple Storage Service) pour entreposer ses données de sondage. Il choisit une région AWS et

définit les paramètres de sécurité appropriés, tels que le chiffrement des données.

4.5.2.1.3 Étape 3 : Transfert des données vers Amazon S3

Le chercheur transfère les données de sondage dans son compartiment Amazon S3 à l'aide de l'interface en ligne AWS ou d'outils d'importation.

4.5.2.1.4 Étape 4 : Configuration des autorisations

Pour sécuriser davantage les données, le chercheur configure les autorisations d'accès aux données dans Amazon S3. Il attribue des rôles et des politiques d'accès spécifiques aux utilisateurs, garantissant que seules les personnes autorisées peuvent accéder aux données.

4.5.2.1.5 Étape 5 : Configuration d'accès dans RStudio

Le chercheur installe le package *aws.s3* dans RStudio pour accéder à ses données entreposées dans Amazon S3. Il configure également les informations d'identification AWS dans son environnement RStudio.

4.5.2.1.6 Étape 6 : Accès et analyse des données dans RStudio

À l'aide du package *aws.s3*, le chercheur peut maintenant accéder à ses données directement dans RStudio par quelques lignes de code. Il peut charger les données dans des structures de données R et effectuer des analyses statistiques, des visualisations et des croisements.

4.5.2.1.7 Étape 7 : Sécurité et conservation des données

Après avoir effectué ses analyses, le chercheur peut choisir de conserver les données de sondage dans Amazon S3 en utilisant les politiques de conservation appropriées. Il peut également archiver des copies de sauvegarde pour garantir l'intégrité des données à long terme.

Dropbox se concentre principalement sur l'entreposage et la collaboration de fichiers, alors que AWS offre une gamme de services *cloud*, y compris l'entreposage sécurisé de données sensibles et la mise en place d'infrastructures évolutives. GitHub, d'autre part, se concentre sur la gestion de versions et la collaboration de code source. Chaque outil a son propre domaine d'expertise et peut être utilisé de manière complémentaire pour différents aspects de la recherche.

4.5.3 Conclusion

L'entreposage des données est une étape cruciale dans la recherche en sciences sociales numériques. Choisissez des outils adaptés à la sensibilité des données, privilégiez des services sécurisés comme AWS pour les données sensibles, et utilisez Dropbox pour la collaboration et l'entreposage de fichiers non sensibles. Une gestion efficace des versions, de la structure des dossiers et de la sécurité garantira l'intégrité de vos données et facilitera la collaboration tout au long de vos projets de recherche.

5 Gestion de la littérature

5.1 Straight up from Jedro Arnaud

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 adap- tées 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 donnée étant utile à l'étude de la littérature désignée. Cette étape est deman- dante, car les chercheur.euse.s doivent se conformer à une grille de codi- fication 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 per- sonnes chercheuses. Les plateformes DistillerSR, Archie et Rayyan sont notamment largement utilisées par les personnes chercheuses.

6 La gestion des références

6.1 Pourquoi citer ?

La citation des sources est une pratique incontournable dans le monde académique, essentielle à la préservation de la crédibilité académique et au maintien des normes éthiques. Elle sert de fondement à la contextualisation de nos recherches, nous permettant de situer nos travaux au sein d'un cadre scientifique établi et reconnu. Ce processus de contextualisation facilite non seulement la compréhension de l'évolution des connaissances dans un domaine donné, mais contribue également à la création d'une base de connaissances solide et dynamique, sur laquelle d'autres travaux peuvent être bâtis (Zaid, Shamsudin, and Habil 2017). La référenciation rigoureuse des travaux antérieurs garantit la reproductibilité des expériences et des analyses, un pilier central de la méthodologie scientifique. En fournissant des détails précis sur les méthodes et résultats, nous ouvrons la voie à la validation et à l'éventuelle réfutation de nos travaux, renforçant ainsi l'intégrité de la recherche (Hughes 2013). De plus, la citation adéquate des sources est une marque de respect envers les contributions des autres chercheurs, assurant une juste attribution du mérite. Cela reconnaît l'importance de chaque découverte et idée dans l'avancement de la science, tout en prévenant le plagiat, une faute grave dans la recherche académique (Racz and Marković 2018).Enfin, une référenciation minutieuse aide à éviter les biais, en exposant clairement les fondements sur lesquels se base notre recherche. Cela permet une évaluation critique des sources et des perspectives, encourageant une approche plus équilibrée et

nuancée dans l'analyse scientifique (Kostoff and Cummings 2013).En résumé, la citation des sources est un acte fondamental qui englobe et adresse de multiples aspects cruciaux de la recherche académique : de la crédibilité et la contextualisation à la reproductibilité, de la création d'une base de connaissances solide à l'attribution correcte du mérite, tout en combattant le plagiat et en minimisant les biais. C'est dans ce contexte que des outils tels que Zotero prennent toute leur importance, en facilitant la gestion rigoureuse des références et en soutenant les chercheurs dans leur quête de rigueur et d'excellence académiques.

6.2 À quoi sert un logiciel de gestion bibliographique ?

Un outil de référence bibliographique est un logiciel conçu pour aider les scientifiques à gérer et à organiser leurs références bibliographiques de manière efficace. Ces outils s'avèrent particulièrement utiles lors de la rédaction d'articles de recherche, de thèses, de mémoires ou d'autres documents académiques. Voici quelques-unes des fonctions principales d'un tel outil :

1. Collecte de références : Les outils de référence bibliographique permettent aux personnes utilisatrices de collecter et d'importer des références bibliographiques à partir de bases de données, de catalogues de bibliothèques, de sites Web ou d'autres sources. Certains outils offrent même la possibilité d'extraire automatiquement les métadonnées à partir de documents PDF.

2. Organisation et classement : Les références collectées peuvent être organisées en différentes catégories et dossiers. Cela facilite la recherche ultérieure et permet de garder une vue d'ensemble claire de la bibliographie.

3. Citation et génération de bibliographies : L'un des avantages majeurs des outils de référence est leur capacité à générer automatiquement des citations et des bibliographies conformes à différents styles de citation (APA, MLA, Chicago, etc.). Ce processus permet de gagner énormément de temps en formatage. Les personnes utilisatrices peuvent insérer des références directement dans leurs documents sans avoir à se soucier des détails de formatage.

4. Collaboration : Certains outils offrent la possibilité de collaborer en ligne, ce qui donne l'occasion à plusieurs personnes de travailler sur une bibliographie commune. Cela peut être utile pour les projets de groupe ou de recherche partagée comme c'est le cas dans une chaire de recherche. En plus d'utiliser un même logiciel, l'utilisation d'un outil de référencement contribue à économiser du temps par la centralisation des données sur un même interface.

5. Recherche et exploration : De nombreux outils de référence bibliographique offrent des fonctionnalités de recherche avancée qui facilitent la découverte de nouvelles références liées à un sujet spécifique.

6. Synchronisation et sauvegarde : Les références et les bibliographies peuvent être synchronisées sur plusieurs appareils, ce offre la possibilité aux personnes utilisatrices d'accéder à leurs références où qu'elles soient. Les sauvegardes régulières assurent que les données ne soient pas perdues en cas de problème technique.

7. Suivi de lecture : Certains outils permettent aux personnes utilisatrices de suivre les articles et les documents qu'elles ont lus, ce qui est particulièrement utile pour garder une trace de la littérature pertinente.

8. Importation et exportation : Les outils de référence bibliographique autorisent généralement l'importation et l'exportation des références dans différents formats, ce qui facilite le transfert de données.

En résumé, un outil de référence bibliographique simplifie grandement le processus de gestion des références bibliographiques, de formatage ainsi que de création de bibliographies. De plus, ces outils offrent la flexibilité de changer de style de citation instantanément, facilitant l'adaptation aux exigences variées des revues scientifiques et permettant aux chercheurs de se consacrer à l'essence de leurs recherches sans se préoccuper des contraintes formelles et des détails de formatage. D'ailleurs, il existe plusieurs outils de référence bibliographique, dont : Endnote, Zotero et Mendeley.

Chaque logiciel offre des caractéristiques uniques tout en partageant des objectifs communs fondamentaux. Ils visent principalement à optimiser l'efficacité et la collaboration en centralisant les références, une commodité indéniable pour les chercheuses et chercheurs et les équipes académiques. Le choix d'un logiciel adapté aux besoins spécifiques des personnes l'utilisants dépend de plusieurs facteurs, notamment la nécessité de partager les résultats de recherche et de collaborer sur des projets communs. Lorsque la collaboration est au cœur d'un projet, il est judicieux que tous les membres de l'équipe adoptent le même outil pour faciliter l'échange d'informations et la cohésion du groupe.

Zotero, EndNote et Mendeley, bien qu'ils partagent des principes de base similaires, se distinguent par des fonctionnalités spécifiques qui peuvent mieux s'aligner sur les préférences et exigences individuelles. La sélection d'un logiciel doit donc être guidée par une évaluation attentive de ses capacités à répondre aux besoins de l'utilisateur, tout en considérant des aspects cruciaux tels que le partage, la collaboration et la facilité d'utilisation.

Il est essentiel de souligner l'importance de la préférence personnelle dans ce choix. L'interface utilisateur, la facilité d'intégration dans les flux de travail existants, et la compatibilité avec d'autres outils numériques sont des critères qui influent grandement sur l'expérience utilisateur et, par conséquent, sur la productivité. En fin de compte, l'outil idéal est celui qui non seulement facilite la gestion des références mais s'intègre de manière transparente dans le quotidien académique de l'utilisateur, lui permettant ainsi de se concentrer pleinement sur la substance de ses recherches.

6.3 Pourquoi Zotero?

L'avantage de Zotero réside dans sa gratuité et son accessibilité libre. Son code est ouvert et son dépôt GitHub compte plus de 13 000 contributions. Il propose une large gamme de fonctionnalités ainsi que la possibilité d'ajouter des extensions, complétant ainsi son usage. Zotero est puissant tout en restant facile à utiliser. Il est disponible sur plusieurs plateformes (Windows, Mac, Linux, iOS, Android), favorisant ainsi la collaboration entre tous les membres d'une équipe de recherche utilisant des plateformes diverses. Il est possible de synchroniser sa bibliothèque Zotero sur plusieurs appareils, soit en utilisant le service cloud payant de Zotero, soit en configurant son propre espace de stockage cloud. Zotero s'intègre parfaitement dans un projet de recherche utilisant LaTeX ou Quarto, car il permet de générer des fichiers .bib à partir des bibliothèques et de les maintenir à jour automatiquement. Il s'intègre également aux logiciels de traitement de texte tels que LibreOffice et Microsoft Office. Il est possible de générer des bibliographies et des citations dans plus de 9 000 styles de citation différents, ce qui le rend adaptable à tous les besoins.

Un autre grand avantage de Zotero réside dans la centralisation des sources bibliographiques et de leurs fichiers associés. Il est possible d'ajouter des PDF à Zotero et de les synchroniser au sein de groupes de travail, facilitant ainsi le partage de documents avec les autres membres de l'équipe de recherche. Plus besoin de recourir à des dossiers partagés ou d'envoyer des documents par courriel ou via des plateformes de partage de fichiers : tout est centralisé dans Zotero. Cette centralisation offre la possibilité d'effectuer des recherches par mot-clé (type ctrl+f) à travers l'ensemble des sources d'une bibliothèque. Vous rédigez une conclusion sur les radis finlandais et souhaitez discuter des enjeux internationaux liés à leur agriculture en citant une source consultée il y a trois ans ? Avec Zotero, retrouver cette source ne prend que quelques secondes grâce à une simple recherche.

L'inconvénient de Zotero réside dans sa difficulté à gérer d'immenses biblio-

thèques contenant plusieurs milliers de fichiers, nécessitant parfois l'achat d'espace de stockage supplémentaire. De plus, bien que performant, le logiciel n'est pas exempt de défauts. Il arrive qu'il faille compléter manuellement des informations non détectées automatiquement par le connecteur intégré.

Zotero est souvent utilisé en combinaison avec BibLaTeX via l'extension Better BibTeX pour exporter et actualiser automatiquement des bibliographies au format .bib. BibLaTeX, une extension moderne pour gérer les bibliographies dans LaTeX et Quarto, s'utilise couramment avec Biber, un outil de traitement bibliographique avancé compatible avec BibLaTeX. Biber propose des fonctionnalités telles que le tri poussé, la gestion de multiples bibliographies et le traitement de divers formats de données bibliographiques. BibLaTeX, prenant en charge de nombreuses langues, est idéal pour la rédaction de documents destinés à un public international. L'exportation de bibliothèques Zotero sous forme de fichiers .bib pour leur utilisation avec BibLaTeX est simplifiée grâce à Better BibTeX, qui assure la mise à jour automatique de ces fichiers. Il est recommandé de maintenir dans votre fichier .bib uniquement les références utilisées, organisées par ordre alphabétique, afin de faciliter la collaboration et le partage des ressources.

6.4 BibLaTeX

BibLaTeX est une extension destinée au traitement des bibliographies dans LaTeX et Quarto, généralement associée à Biber, un programme conçu pour le traitement des données bibliographiques spécifiquement pour BibLaTeX. Biber propose des fonctionnalités avancées comme le tri poussé, la gestion de multiples bibliographies et la capacité de traiter des sources bibliographiques dans divers formats. Grâce à sa prise en charge étendue des langues, BibLaTeX est particulièrement adapté à la rédaction d'articles ou de livres pour un public international. Bien que BibLaTeX soit avant tout un package pour LaTeX, il est possible d'exporter des bibliothèques

depuis des outils tels que Zotero sous forme de fichiers .bib, qui peuvent ensuite être exploités avec BibLaTeX. L'extension Better BibTeX permet de maintenir automatiquement à jour vos fichiers .bib à partir de Zotero. Il est recommandé de conserver uniquement les références utilisées dans votre fichier et de les organiser par ordre alphabétique, facilitant ainsi la coopération et le partage des sources.

Les inconvénients potentiels de BibLaTeX comprennent une courbe d'apprentissage plus accentuée pour ceux habitués à BibTeX, ainsi que le besoin de mises à jour régulières pour assurer la compatibilité avec les versions les plus récentes de LaTeX. En outre, certains éditeurs académiques ou revues possèdent leurs propres styles de citation et peuvent ne pas accepter les soumissions réalisées avec BibLaTeX, même si cette réticence tend à diminuer.

En conclusion, pour ceux qui cherchent à maximiser la flexibilité et la puissance de leurs outils de gestion de bibliographie dans LaTeX, BibLaTeX, en tandem avec Biber, offre une solution moderne et robuste.

6.5 Installation et configuration de Zotero

Dans cette section, vous serez amené notamment à installer Zotero ainsi que Better BibTeX. Better BibTeX est une extension de Zotero servant à générer et à maintenir à jour des fichiers .bib compatibles avec BibLaTeX, à partir de Zotero.

6.5.1 Zotero

- Installer Zotero
- Installer Zotero Connector

- Une fois Zotero installé, vous avez l'option de créer un compte Zotero. L'identifiant que vous utiliserez sera celui que vous partagerez à vos collaborateurs pour créer et joindre des groupes.

6.5.1.1 Better Bibtex

- La prochaine étape sera d'installer Better BibTex. Pour ce faire, allez dans l'onglet tools > Add-ons ensuite cliquez sur l'icone de paramètre et faites Install Add-on From File. Sélectionnez le fichier .xpi que vous avez téléchargé.

IMPORTANT

- Une fois le module complémentaire installé, accédez aux paramètres de Better BibTeX en allant dans l'onglet Zotero > Préférences > Onglet Better BibTeX > Ouvrir les préférences de Better BibTeX.

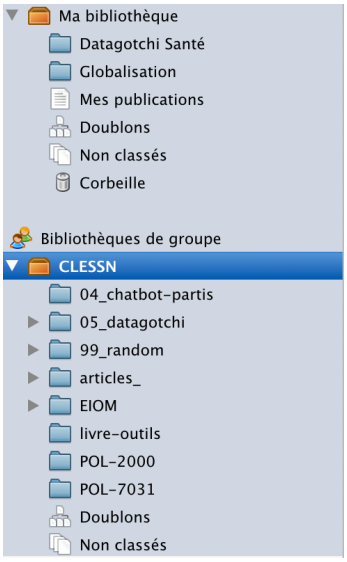
- Il est important, lors de la collaboration, de s'assurer d'avoir les mêmes clés de citation que vos collègues. Better BibTeX peut s'assurer que vos clés respectent un format standard.

- Voici une suggestion de format de clé de citation : il s'agit simplement du nom de l'auteur et de l'année de publication à deux chiffres. Pour l'utiliser, collez ceci dans la section Format de clé de citation : `authEtal2.fold.lower.replace(find=".",replace=_) + len + shortyear | veryshorttitle + shortyear`

- Afin de vous assurer d'avoir les mêmes clés de citation, vous pouvez faire un clic droit sur vos références, aller dans les options de Better BibTeX et cliquer sur "Actualiser les clés de citation"

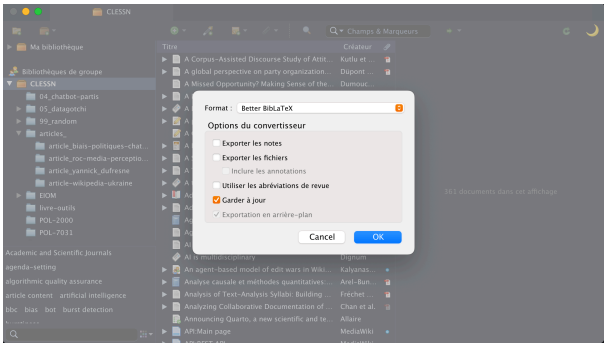
6.5.1.2 Génération du fichier .bib

Dans Zotero, vous devriez maintenant voir le groupe Zotero de votre équipe dans les Group Libraries.



Il est important de comprendre que tout changement que vous faites dans Zotero sera automatiquement synchronisé avec le groupe de votre équipe de travail. Ainsi, si vous supprimez une référence, elle sera supprimée pour tout le monde!

Clic-droit sur la collection livre-outils > Export Collection choisissez le format Better BibLaTex et cochez la case [x] Keep updated. Faites OK et sauvegardez le fichier dans le dossier .git du projet livre-outils. Ce dossier sera constamment mis à jour avec les changements que vous faites dans Zotero et sera synchronisé avec le projet Github quand vous ferez vos pull requests.



6.5.1.3 Utilisation de Zotero lors de l'écriture

Lors de l'écriture, vous n'avez qu'a écrire @ dans votre éditeur pour faire sortir la palette de référencement.

6.5.1.4 Ajouter des références à Zotero

Il y a différentes façons d'ajouter des références à Zotero :

- Glisser-déposer à partir de votre bibliothèque personnelle.
- Glisser-déposer les PDF que vous avez sur votre ordinateur dans la collection Livres-Outils. Zotero va essayer de trouver les métadonnées automatiquement.
- Si cela ne réussit pas, vous pourrez ajouter la référence en cliquant sur la baguette magique en haut à gauche du symbole "+" vert. L'outil de la baguette magique est utile si vous possédez le DOI ou l'ISBN de l'article/livre que vous devez ajouter. Dans les rares cas où Zotero ne trouve rien concernant votre référence, vous pourrez remplir les différents champs manuellement.
- Utiliser le connecteur dans votre navigateur. Zotero tentera également de télécharger l'article directement et de l'inclure dans la collection appropriée.

6.6 Conclusion

Pour conclure, l'adoption de Zotero comme outil de gestion bibliographique se révèle être un choix judicieux pour tout chercheur soucieux de l'efficacité et de la rigueur dans le processus de documentation scientifique. Au-delà de la simple facilitation du travail de recherche en équipe, Zotero se distingue par sa capacité à optimiser la gestion des citations et des bibliographies, permettant ainsi une économie de temps considérable et une réduction des risques d'erreurs. Sa fonctionnalité de centralisation des sources et de leurs fichiers associés offre un avantage notable en termes d'organisation et d'accès rapide à l'information, cruciale dans le cadre de recherches approfondies ou pluridisciplinaires. L'intégration de Zotero dans les environnements académiques, même en dehors des contextes de recherche, comme l'enregistrement des lectures pour des cours ou des séminaires, prépare efficacement les utilisateurs à des pratiques de recherche plus poussées et renforce la culture de la gestion rigoureuse des références. Cette initiation précoce est d'autant plus pertinente que Zotero se prête à une variété de styles de citation, répondant ainsi aux exigences diverses des publications académiques. Il est important de souligner que la maîtrise de Zotero, bien que facilitée par de nombreux tutoriels et ressources en ligne, représente un investissement en temps qui se trouve largement compensé par les bénéfices en termes d'efficacité et de qualité du travail de recherche. En outre, l'accès gratuit et le caractère open-source de Zotero témoignent de son engagement en faveur d'une diffusion élargie du savoir et d'une collaboration scientifique ouverte.

6.6.1 Revue Systématique

une revue systématique est une synthèse méthodique et exhaustive de la littérature centrée sur une question de recherche spécifique. Elle vise à recenser l'ensemble des travaux académiques, publiés ou non, relatifs à cette question Kibbee (2023) **kibbee23**. Comme mentionné précédemment, une telle revue requiert un investissement conséquent en temps et