

# **Outils de recherche en sciences sociales numériques**

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

2023-08-26



# Table of contents

|  |          |
|--|----------|
| <b>Avant-propos</b>  | <b>1</b> |
| <b>Introduction</b>  | <b>3</b> |
| <b>1 Trois défis pour une contribution aux sciences sociales numériques</b>  | <b>5</b> |
| 1.1 Défi #1: l'inévitable défi technique. . . . .  | 6        |
| 1.2 Défi #2: le nécessaire défi théorique. . . . .   | 8        |
| 1.3 Défi #3: l'épineux défi éthique. . . . .   | 8        |
| 1.4 Conclusion de cette première partie du chapitre . . . . .  | 9        |
| 1.5 Comment les données massives affectent-elles les sciences sociales? Changements actuels et quelques réflexions sur l'avenir . . . . .  | 9        |
| 1.6 Définition des données massives . . . . .  | 11       |
| 1.7 Les données massives et les sciences sociales . . . . .  | 12       |
| 1.8 La validité de la mesure en sciences sociales . . . . .  | 13       |
| 1.9 La validité des données massives . . . . .   | 14       |
| 1.9.1 Validité interne des données massives . . . . .  | 15       |
| 1.9.2 Validité externe des données massives . . . . .  | 16       |
| 1.9.3 Données expérimentales . . . . .   | 17       |
| 1.9.4 Données observationnelles . . . . .  | 18       |
| 1.9.5 Validité écologique et observation par sous-groupes .  | 19       |
| 1.10 Vers le futur : les données massives effectueront-elles un changement dans la posture épistémologique en sciences sociales? . . . . . | 20       |

## Table of contents

|          |  |           |
|----------|--|-----------|
| <b>2</b> | <b>Le monde du libre</b>   | <b>21</b> |
| 2.1      | Émergence et ascension . . . . .   | 22        |
| 2.1.1    | Logiciel libre et <i>open source</i> . . . . .                                   | 23        |
| 2.2      | Principaux avantages et inconvénients . . . . .                                  | 23        |
| <b>3</b> | <b>Les outils de collecte de données</b>   | <b>25</b> |
| 3.1      | <b>Le Big Data et les différents acteurs de la société : .</b>                   | <b>26</b> |
| 3.2      | Les outils de collecte de données de sondages . . . . .                          | 27        |
| 3.3      | <b>Factiva : outils de récolte de données médiatiques .</b>                      | <b>27</b> |
| 3.4      | <b>Les extracteurs : avoir accès à des données massives via du code. . . . .</b> | <b>31</b> |
| 3.5      | <b>Covidence : outil de récolte d'articles scientifiques .</b>                   | <b>32</b> |
| 3.6      | Conclusion et discussion: . . . . .  | 34        |
| <b>4</b> | <b>R ou ne pas R?</b>  | <b>35</b> |
| 4.1      | Pourquoi R? . . . . .  | 35        |
| 4.2      | Trucs et astuces . . . . .   | 37        |
| 4.3      | Les environnements de développement intégré . . . . .                            | 38        |
| 4.4      | Où coder en R ? . . . . .  | 38        |
| 4.5      | Pourquoi RStudio ? . . . . .   | 39        |
| 4.5.1    | Qu'est-ce que RStudio ? . . . . .  | 39        |
| 4.5.2    | Avantages et inconvénients de RStudio . . . . .                                  | 40        |
| 4.6      | Comment utiliser RStudio ? . . . . .   | 42        |
| 4.7      | Personnaliser son RStudio . . . . .  | 44        |
| <b>5</b> | <b>Baliser les sciences sociales : langages et pratiques</b>                     | <b>45</b> |
| 5.1      | Réflexion théorique . . . . .  | 49        |
| 5.2      | Réflexion méthodologique . . . . .   | 58        |
| 5.3      | Trucs et astuces . . . . .   | 61        |
| 5.3.1    | Logiciels de bureau . . . . .  | 62        |
| 5.3.2    | Logiciels en ligne . . . . .   | 62        |
| 5.4      | Références . . . . .   | 62        |

|           |  |           |
|-----------|--|-----------|
| <b>6</b>  | <b>La gestion des références</b>   | <b>63</b> |
| 6.1       | <u>Pourquoi citer ?</u> . . . . .  | 63        |
| 6.2       | <u>À quoi sert un logiciel de gestion bibliographique ?</u> . . . . .  | 63        |
| 6.3       | Installation et configuration de Zotero . . . . .  | 68        |
| 6.3.1     | Zotero . . . . .   | 68        |
| <b>7</b>  | <b>Une image vaut mille mots</b>   | <b>71</b> |
| 7.1       | Introduction . . . . .   | 71        |
| 7.2       | Réflexion théorique . . . . .  | 72        |
| 7.2.1     | Les options disponibles . . . . .  | 72        |
| 7.3       | Références . . . . .   | 78        |
| <b>8</b>  | <b>Outils de recherche: la quête infinie de l'optimisation</b>   | <b>79</b> |
| 8.1       | Introduction du chapitre . . . . .   | 79        |
| 8.2       | Logiciel de gestion de communication (Slack) . . . . .   | 80        |
| 8.3       | Logiciel de gestion de projet . . . . .  | 81        |
| 8.4       | Stockage de données (Hub, AWS, Dropbox == défi éthique)  | 81        |
| 8.5       | Logiciel de gestion de versions décentralisé . . . . .   | 81        |
| 8.6       | Conclusion: un exemple de l'utilisation des outils . . . . .   | 82        |
| <b>9</b>  | <b>Outils d'intelligence artificielle</b>  | <b>83</b> |
| <b>10</b> | <b>Sapientia Artificiosa: L'Excellence Cognitive au Service des Études Sociales"</b>   | <b>85</b> |
| <b>11</b> | <b>En général, c'est quoi l'IA, le ML, le DL, etc.? (en mentionnant que c'est un champ qui change tout le temps) (JOZEF)</b> | <b>87</b> |
| <b>12</b> | <b>La révolution ChatGPT</b>   | <b>89</b> |
| <b>13</b> | <b>Deepl, google traduction, etc,</b>  | <b>91</b> |
| <b>14</b> | <b>Midjourney, génération d'images, etc.</b>   | <b>93</b> |
| <b>15</b> | <b>1 autre exemple?</b>  | <b>95</b> |

*Table of contents*

|   |            |
|---|------------|
| <b>16 Machine Learning, reconnaissance d'images etc</b>       | <b>97</b>  |
| <b>17 LLM, chatbots, etc.</b>                                 | <b>99</b>  |
| <b>18 Serpents et échelles</b>                                | <b>101</b> |
| 18.1 Introduction . . . . .                                   | 101        |
| 18.2 Débutant . . . . .                                       | 102        |
| 18.2.1 Environnements de programmation . . . . .              | 102        |
| 18.2.2 Les alternatives à Word : les langages de balisage . . | 102        |
| 18.2.3 Serpents . . . . .                                     | 102        |
| 18.3 Intermédiaire . . . . .                                  | 103        |
| 18.3.1 La gestion des références . . . . .                    | 103        |
| 18.3.2 Visualisation graphique en R . . . . .                 | 103        |
| 18.3.3 Serpents : . . . . .                                   | 103        |
| 18.4 Avancé . . . . .   | 105        |
| 18.4.1 Gestion de projet et de données . . . . .              | 105        |
| 18.4.2 Outils d'intelligence artificielle . . . . .           | 105        |
| 18.4.3 Snakes . . . . .                                       | 105        |
| 18.5 Conclusion . . . . .                                     | 105        |
| <b>References</b>   | <b>107</b> |

# **Avant-propos**

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





# Introduction



# 1 Trois défis pour une contribution aux sciences sociales numériques

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

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

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

## *1 Trois défis pour une contribution aux sciences sociales numériques*

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

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

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

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

### **1.1 Défi #1: l'inévitable défi technique.**

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

### 1.1 Défi #1: l'inévitable défi technique.

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

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

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

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

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

## *1 Trois défis pour une contribution aux sciences sociales numériques*

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

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

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

### **1.2 Défi #2: le nécessaire défi théorique.**

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

### **1.3 Défi #3: l'épineux défi éthique.**

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

### 1.4 Conclusion de cette première partie du chapitre

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

## 1.4 Conclusion de cette première partie du chapitre

- Au travers des nouveaux apprentissages et des exemples qui sont offerts dans ce livre, le lecteur est encouragé à se poser ces 3 questions:
  1. D’abord, comment puis-je utiliser ces nouveaux outils pour faire progresser les méthodologies de recherche actuelles?
  2. Ensuite, comment puis-je utiliser ces nouveaux outils pour contribuer à l’avancement des théories de mes champs de recherche?
  3. Enfin, comment puis-je utiliser ces nouveaux outils pour exercer un impact positif sur mes semblables?

## 1.5 Comment les données massives affectent-elles les sciences sociales? Changements actuels et quelques réflexions sur l’avenir

L’apparition des données massives (*big data*) dans le paysage technologique représente un de ces cas de plus en plus commun de phénomène hautement

## 1 Trois défis pour une contribution aux sciences sociales numériques

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

Dans le domaine des sciences sociales, les avancées dûes à l'utilisation des données massives se font de plus en plus fréquentes et l'impact des données massives dans le domaine de la recherche sociale est en ce sens indéniable. Toutefois, d'un point de vue épistémologique, l'utilisation des données massives en recherche en sciences sociales dans les dernières années laisse plusieurs questions ouvertes dans son sillage.

Comment l'utilisation des données massives change-t-elle la pratique des sciences sociales? Les données massives causeront-elles un changement de paradigme scientifique? Quels impacts auront-elles sur les traditions scientifiques dominantes (e.g., béhavioralisme, individualisme méthodologique) en sciences sociales?

Ce chapitre ne prétend pas offrir de réponses définitives à ces questions, mais plutôt des pistes de réflexion par le biais d'une introduction critique à certains points relatifs aux impacts des données massives sur la recherche en sciences sociales. Premièrement, je présente une conceptualisation des données massives. Deuxièmement, je me penche sur les impacts des données massives en sciences sociales et souligne tout particulièrement comment elles affectent les enjeux de la *validité* interne et externe dans la



## 1.6 Définition des données massives

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

### 1.6 Définition des données massives

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

1. Premièrement, les données massives représentent une (1) ***quantité importante de points d’information*** qui varient selon la nature, le type, la source, etc. En ce sens, la distinction est simplement quantitative. Il s’agit d’une première dimension à la définition des données massives.
2. Deuxièmement, d’une perspective technique et technologique, les données massives constituent un (2) ensemble de ***pratiques*** de collecte, de traitement et d’analyse de ces points d’information. Les données massives représentent donc une technique ou une méthode nouvelle de recherche.
3. Finalement, d’une perspective sociologique, les données massives représentent (3) un phénomène incorporant à la fois la dimension propre aux ***développements technologiques, ainsi que les impacts sociétaux de ces développements*** – i.e., les risques à la confidentialité des données, les enjeux relatifs au consentement et à l’autorisation de collecte des informations, les innovations en intelligence artificielle, etc. Cette perspective souligne le caractère essentiellement social des données massives.

Dans les domaines scientifiques et technologiques, la définition courante donnée aux données massives intègre des éléments de ces trois niveaux

## 1 Trois défis pour une contribution aux sciences sociales numériques

d'analyses en se référant à la composition et à la fonction des données. Premièrement, la *composition* des données massives est généralement conceptualisée comme comprenant « 4V » : le volume, la variété, la vélocité et la véracité. Cette conceptualisation jouie d'un large consensus scientifique (Chen, Mao et Liu, 2014; Gandomi et Haider, 2015; Kitchin et McArdle 2016).

Par ailleurs, plusieurs chercheurs ont élargi cette définition de la composition des données massives en y incluant, par exemple, la variabilité et la valeur des points de données (CITE). Deuxièmement, la *fonction* des données massives comprend les innovations relatives à l'optimisation, à la prise de décision et à l'approfondissement des connaissances qui résultent de leur utilisation. Ces fonctions touchent des domaines sociaux disparates, incluant le souci d'efficacité et de rendement du secteur privé et public ainsi que la recherche scientifique pure (Gartner 2012).

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

### 1.7 Les données massives et les sciences sociales

Dans le domaine des sciences sociales, les changements causés par l'utilisation des données massives en recherche sont significatifs. Plusieurs n'hésitent d'ailleurs pas à les qualifier de changement de paradigme

## 1.8 La validité de la mesure en sciences sociales

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

## 1.8 La validité de la mesure en sciences sociales

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

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

Autrement dit, la mesure et les données collectées ne représentent pas le phénomène – l’opinion publique – en soi.

On a déjà mentionné que la validité de la mesure a de l’importance puisqu’elle garantit que ce qui est mesuré représente réellement ce qu’on croit mesurer. Mais pour être plus spécifique, dans une approche positiviste, la validité de la mesure se traduit généralement par une logique de classification des valeurs attribuées aux différentes manifestations distinctes d’un même phénomène. Par exemple, une mesure de la démocratie comme celle proposée par *Freedom House*, fréquemment utilisée en science politique, classe les libertés civiles et les droits politiques des états du monde par degré, de 1 à 7, afin de construire un index allant d’autoritarisme complet à démocratie parfaite. Les scores représentent, dans ce contexte, une mesure artificielle, mais ordonnée et logique, des idées contenues dans le concept de démocratie telles que libertés civiles et droits politiques. On peut ainsi dire que le souci avec la validité de la mesure traverse les connexions entre (1) le phénomène social étudié (la démocratie), (2) son opérationnalisation (via les libertés civiles et droits politiques) et (3) la méthode de mesure utilisée pour observer et classer d’une certaine façon le phénomène et les données qui en découlent (dans le cas de *Freedom House* des codeurs indépendants).

## 1.9 La validité des données massives

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

L’étude de King et al. (2013) représente un cas éclairant de phénomène social que que l’utilisation des données massives a rendu possible d’étudier.

## 1.9 La validité des données massives

En se basant sur la collecte de plus de 11 millions de publications sur les réseaux sociaux chinois, King et al. ont pu mesurer la censure exercée par le gouvernement chinois sur les réseaux sociaux. En utilisant des données massives nouvelles, King et al. ont donc pu observer une manifestation inédite de censure massive qui, sans de telles données, serait probablement demeurer mal comprise d’une perspective scientifique. Le nombre de recherches basées sur l’utilisation des données massives similairement innovantes en sciences sociales est par ailleurs en croissance constante (Beauchamp 2017; Bond et al. 2012; Poirier et al. 2020).

Cependant, il faut aussi souligner que les données massives, de par leur complexité, peuvent avoir pour désavantage d’embrouiller l’étude des phénomènes sociaux. Les opportunités scientifiques liées aux données massives s’accompagnent en effet de certaines difficultés méthodologiques.

Aux nombres de ces difficultés, trois questions sont particulièrement cruciales : (1) la validité interne, (2) la validité externe, et (3) la question d’un changement de posture ou d’orientation épistémologique en sciences sociales causé par les données massives.

### 1.9.1 Validité interne des données massives

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

Un des problèmes liés à cette situation consiste en ce qu’il est difficile de garantir une source *exogène* de variation par laquelle les chercheur.e.s

éliminent l'effet potentiel des facteurs confondants (*confounders*). La distribution aléatoire d'un traitement et d'un contrôle (dans une expérience en laboratoire ou sur le terrain) représente le standard le plus élevé permettant de fournir cette source exogène de variation.

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

Cependant, dans l'état actuel des choses, le volume et la variété (2 des 4 V) des données massives (textuelles, numériques, vidéos, etc.) peuvent miner la qualité de l'inférence causal entre 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.

### 1.9.2 Validité externe des données massives

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

Le cas célèbre de l'erreur de prédiction du *Literary Digest* lors de la campagne présidentielle américaine de 1936 illustre bien ce problème récurrent. Le *Literary Digest* a effectivement prédit à tort la victoire de Alf Landon, le candidat du parti républicain, sur Franklin D. Roosevelt, le candidat démocrate, parce que l'échantillon de répondants utilisé par le *Literary*

## 1.9 La validité des données massives

*Digest* dans son sondage a surreprésenté 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%.

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

### 1.9.3 Données expérimentales

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

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

#### 1.9.4 Données observationnelles

En ce qui à trait aux données observationnelles, il y a deux points importants. Premièrement, des méthodes d'inférence basées sur des approches par « design » (*design-based methods*) comme une méthode de régression sur discontinuité, de variable instrumentale, etc. peuvent également garantir des inférences explicatives et causales valides. Elles nécessitent toutefois plusieurs postulats plus restrictifs dont l'objectif est d'imiter ou de recréer, de la manière la plus fidèle possible, une distribution aléatoire du traitement – ce que la littérature appelle un « *as-if random assignment* » (Dunning, 2008).

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

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

Dans un tel cas, des méthodes de pondération sur des données observationnelles peuvent compenser pour la sur- ou la sous-représentativité de sous-groupes dans un échantillon afin d'assurer la validité de l'inférence



## 1.9 La validité des données massives

entre échantillon et population. Les données massives ont ici une importance puisqu'une pondération fiable nécessite une quantité substantielle d'observations. Une pondération *a posteriori* sera donc plus fiable plus l'échantillon est grand. Les données massives ont ainsi une valeur ajoutée afin d'établir des inférences descriptives plus précises et sophistiquées.

### 1.9.5 Validité écologique et observation par sous-groupes

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

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

## 1 Trois défis pour une contribution aux sciences sociales numériques

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

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

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

## 2 Le monde du libre

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

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

---

<sup>1</sup>La redistribution doit évidemment respecter certaines conditions précises, dont l'enfreint peut mener à des condamnations

## 2 Le monde du libre

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

### 2.1 Émergence et ascension

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

---

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

<sup>2</sup>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.

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

## 2.2 Principaux avantages et inconvénients

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

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

### 2.1.1 Logiciel libre et *open source*

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

## 2.2 Principaux avantages et inconvénients

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

## *2 Le monde du libre*

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

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

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

### 3 Les outils de collecte de données

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

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

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

Chacun de ces champs de recherches se voit confronté à une panoplie de défis théoriques et techniques en lien avec l'émergence des données massives. La révolution technologique permet une étude plus approfondie des phénomènes auxquels sont confrontés les différents acteurs de la société démocratique. Toutefois, la collecte de données permettant de mener à termes de telles études peut s'avérer complexe. Pour chaque pilier de la démocratie, les sections suivantes énumèrent et expliquent les capacités techniques d'outils permettant aux chercheurs d'accéder à des données massives. Bien que d'autres outils existent et offrent des résultats satisfaisants, les méthodes suivantes sont particulièrement pertinentes dans une optique d'étude des sciences sociales numériques.



## **3.2 Les outils de collecte de données de sondages**

### **3.3 Factiva : outils de récolte de données médiatiques**

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

L'outil Factiva offre un accès à l'ensemble des articles d'une panoplie de médias provenant d'une vaste sélection de pays. Le moteur de recherche est opéré par Dow Jones et offre également l'accès à des documents d'entreprises. Toutefois, l'accès qu'il offre aux contenus de média est particulièrement pertinent pour la communauté scientifique en communication et en sciences sociales. Il offre accès à plus de 15 000 sources médiatiques provenant de 120 pays. Il permet de télécharger une quantité illimitée de documents RTF pouvant contenir jusqu'à 100 articles médiatiques chacun. Les articles peuvent être sélectionnés automatiquement en cochant le bouton proposant de sélectionner tous les

### 3 Les outils de collecte de données

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

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

#### 3.3.0.0.1 Figure 1. Interface de recherche de Factiva

### 3.3 Factiva : outils de récolte de données médiatiques

Recherche de texte libre

Recherche Guidée

Exemples

Date: **Au cours des 3 derniers mois** Doubtons: **Désactivé** **Rechercher**

Source: Toutes les sources

Auteur: Tous les auteurs

Société: Toutes les sociétés

Recherche des Experts de Factiva

Sujet: Tous les sujets

Secteur économique: Tous les secteurs économiques

Région: Toutes les régions

Chercher

Langue: Toutes les langues

Plus d'options

Effacer la recherche **Rechercher**

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

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

#### 3.3.0.0.2 Figure 2. Interface de recherche d'Eureka

### 3 Les outils de collecte de données

The screenshot displays the EUREKA search platform's advanced search interface. At the top, a navigation bar includes the EUREKA logo, a 'RECHERCHER' button, and links for 'DOSSIERS' and 'PUBLICATIONS PDF'. The main heading is '< Recherche avancée'. Below this, there's a section for 'Mots clés dans tout le texte' with a large text input field. To the right of this field are three rows of search filters, each with a red 'ET' button, a grey 'OU' button, and a grey 'SANS' button, followed by a dropdown menu. The first row is set to 'dans le titre', the second to 'dans l'introduction', and the third to 'dans le nom de l'auteur'. Below these filters is a button labeled 'Ajouter une zone de mots clés'. The 'Sources' section follows, with a prompt 'Sélectionnez vos sources par :'. Three radio buttons are present: 'groupe de sources' (selected), 'critères de sources', and 'nom de source'. Below this is a 'Domaine de recherche' dropdown menu currently set to 'Tout le contenu'. At the bottom left, the 'Période' section has a dropdown menu set to 'Depuis 30 jours'. On the right side of the interface, there's a 'Astuces de recherche' (Search tips) sidebar. It contains several examples of search queries with their corresponding results: 'pomme verte' (contains the exact phrase), 'blanc & noir' (contains both at once), 'rouge | vert' (contains either), 'pomme & (verte | rouge)' (contains 'pomme' followed by either 'verte' or 'rouge'), 'bières ! "bières blondes"' (contains 'bières' but not 'bières blondes'), 'voiture \$2 sport' (contains 'voiture' followed by 'sport' with a maximum of two words in between), and 'automobile %2 salon' (contains 'automobile' followed by 'salon' with a maximum of two words in between). At the bottom right, there are buttons for 'Recommencer' and 'Recherche'.

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

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

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

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

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

Les données en lien avec les décideurs sont souvent accessibles sur des sites gouvernementaux. Toutefois, certaines identifications peuvent être nécessaires et l'accès peut être compliqué, particulièrement dans une perspective de données massives. C'est dans cette optique que les extracteurs peuvent être utiles. Un code peut extraire de manière automatique les débats des Assemblées nationales, les communiqués de presses des gouvernants, les plateformes électorales des partis politiques, ce qui offre un accès inégalé aux personnes chercheuses aux données de décideurs. Dans une autre optique, des extracteurs peuvent également offrir accès aux données provenant des médias socionumériques comme Twitter ou Facebook, sur lesquels les acteurs politiques sont souvent très actifs. Un extracteur peut, par exemple, être en mesure de répertorier l'ensemble des Tweets de journalistes, de politiciens ou encore de citoyens de manière automatisée, offrant encore une fois un accès inégalé aux personnes chercheuses à des données massives exclusives. L'élaboration d'extracteurs est toutefois facilitée par l'existence d'API sur les plateformes exploitées. Par exemple, Twitter possédait avant les changements de directions récents un API qui facilite l'élaboration d'un scraper. En contrepartie, Facebook ne possède pas d'API, ce qui rend l'accès à ses données beaucoup plus complexe. Un extracteur peut également offrir l'accès à des données médiatiques, en codant un accès à des fils RSS ou encore aux HTML des médias extraits.

### 3.5 Covidence : outil de récolte d'articles scientifiques

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

La plateforme en ligne Covidence est utilisée pour faciliter les revues systématiques de littérature, et cette dernière permet de réduire drastiquement le temps d'accomplissement du travail en plus de le rendre simple et intuitif. L'outil a été développé pour mieux gérer et organiser l'évaluation de quantité importante d'études scientifiques. L'exécution d'une revue de littérature sur Covidence se fait par le biais d'un double codage. C'est-à-dire que l'évaluation des études se fait manuellement par deux codeurs travaillant de manière autonome qui mettront en commun leurs résultats à la fin de l'exercice. L'outil est reconnu pour ses trois étapes précises : « Title and abstract screening », « Full text review » et « Extraction ». Covidence permet d'importer des données massives provenant de base de données bibliographiques. En effet, l'outil lance des requêtes auprès de multiples bibliothèques, ce qui offre l'accès à des milliers d'études sur le champ étudié par les personnes chercheuses. Ces requêtes sont adaptées aux besoins spécifiques de la personne chercheuse voulant explorer en profondeur un domaine de la littérature scientifique.

La première étape, soit le « Title and abstract screening », consiste en la révision des titres et des résumés des articles récoltés. Pour rendre le travail davantage efficace, il est nécessaire d'inclure des critères précis pour

### ***3.5 Covidence : outil de récolte d'articles scientifiques***

analyser les titres et résumés d'articles. En se servant du jugement et des critères qui étaient recherchés, les individus doivent éliminer ou accepter selon la pertinence de l'article quant à la littérature étudiée. Cette partie est souvent longue puisque la littérature existante est souvent massive. Il est donc important pour les personnes chercheuses de se rencontrer à maintes reprises pour discuter des conflits de jugement et pour trouver des compromis. Cette étape, plutôt longue, s'avère très utile et motivante, puisqu'il est possible de développer un jugement critique davantage raffiné et de s'instruire dans une littérature continuellement plus précise.

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

La dernière étape, soit celle de l'extraction, consiste à recueillir toute donnée étant utile à l'étude de la littérature désignée. Cette étape est demandante, car les chercheurs doivent se conformer à une grille codification prédéfinie. Le but est qu'un consensus entre les chercheurs émerge de ce processus. L'extraction permet de faire ressortir les théories, les méthodologies et les conclusions présentent dans les études retenues.

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

### *3 Les outils de collecte de données*

Les plateformes DistillerSR, Archie et Rayyan sont notamment largement utilisées par les personnes chercheuses.

## **3.6 Conclusion et discussion:**

Le précédent chapitre portait sur les différents outils de collecte de données massives mis à la disposition des chercheur.euse.s s'intéressant au champ des sciences sociales numériques. les outils relevés se démarquent par leur capacité de permettre l'accès à des données permettant d'étudier les trois principaux acteurs de la société démocratique, soit les citoyens, les décideurs et les médias. Tel que mentionné à plusieurs reprises lors du chapitre, le but de ce dernier n'est pas d'offrir une liste complète des outils disponibles. Toutefois, les outils énumérés ont été sélectionnés en raison de leur intuitivité, leur relative simplicité d'accès de même que leurs capacités techniques considérées par les auteurs comme étant particulièrement pertinente dans une optique de recherche en sciences sociales numérique.



## 4 R ou ne pas R?

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

### 4.1 Pourquoi R?

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, des choix qui font aujourd'hui la popularité de R auprès d'un large

#### 4 *R* ou ne pas *R*?

pan de la communauté académique. En effet, Morandat et al. (2012) rapportent que le langage a été élaboré afin qu'il soit intuitif et qu'il permette aux nouveaux utilisateurs de rapidement réaliser des analyses.

Le langage de programmation **R** a plusieurs avantages qui le rendent un outil puissant et utile pour tout chercheur. C'est notamment un langage *OpenSource*. Le chapitre précédent ayant déjà élaboré sur le sujet du logiciel libre, il ne sera ici que rappelé les grandes lignes de l'argument, à savoir que : 1) l'*OpenSource* est gratuit d'utilisation; 2) l'*OpenSource* est développé de façon bottom-up, ce qui lui procure une grande flexibilité; et 3) il permet aux utilisateurs de créer leurs propres fonctions. À l'inverse, les logiciels à licences sont coûteux, rigides et l'ajout de fonctionnalités se fait par les développeurs internes à la compagnie ce qui rend le processus plus lent et réduit l'éventail des possibilités. Ceci étant dit, certains avanceront que le c'est justement ce processus interne lent qui assure la validité et la fiabilité des analyses effectuées par SAS, STATA ou SPSS. Or, dans son livre dédié aux utilisateurs de SPSS et de SAS, Muenchen (2011) soulève le point que bien souvent, ce sont des individus atomisés qui développent les nouvelles fonctionnalités de ces langages et que le processus de révisions se fait ensuite par des comités internes de testeurs. Il en va de même pour le développement des *packages* *R* dans la mesure où ce dernier se 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 publie 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.

L'utilisation du langage de programmation **R** peut être perçue

comme ayant quelques inconvénients. Certains disent que la courbe d'apprentissage peut être plus grande que celle de programmes à licences. La véracité de cet argument est discutable. Les programmes demandant des licences ont également un coût d'entrée et les nouvelles itérations des logiciels amènent des changements demandant une période d'adaptation. 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 participer à la communauté de programmation **R**. Toutefois, 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 et les *packages* doivent être suffisamment documentés. Il est également possible d'y télécharger le langage de programmation **R**. Ce dernier, ainsi que ces différents *packages*, est disponible sur Windows, macOS et Linux.

## 4.2 Trucs et astuces

- 10 choses à garder en tête lorsque l'on apprend R :
  1. Vous n'allez pas briser votre ordinateur.
  2. C'est en "gossant" que l'on apprend! Essayez des trucs, expérimentez – souvenez-vous de 1.
  3. Contrairement à la vraie vie, il y a toujours le `ctrl-z` pour vous sauver.
  4. Ayez de l'empathie pour vos futurs lecteurs, ou du moins pour le futur vous – COMMENTEZ VOTRE CODE!
  5. Même Wozniak était mauvais au début.
  6. Pensez à votre santé – levez-vous de votre chaise aux 30 minutes.
  7. S'il y a un bogue, et il y en aura, *Google* est votre ami.

#### 4 *R* ou ne pas *R*?

8. Souvent c'est une question de type de variable – caractère, numérique, facteurs, etc.
9. Si vous faites beaucoup de copier-coller de code, il y a sûrement une façon de l'automatiser.
10. Sérieusement, faites le 4!

### 4.3 Les environnements de développement intégré

#### 4.4 Où coder en *R* ?

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

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

*R* est un des langages de statistiques et d'exploration de données les plus populaires en sciences sociales et il est open-source. Par conséquent, il est logique de choisir un environnement de programmation open-source. *R* est pris en charge par de nombreux environnements de programmation. Plusieurs ont été spécialement conçus pour la programmation en *R* – le plus notable étant RStudio – tandis que d'autres sont des environnements de programmation universels – tel que Visual Studio – et prennent en

## 4.5 Pourquoi RStudio ?

charge R via des plugins. Il est également possible de coder en R à partir d'une interface en ligne de commande. Une telle méthode permet la communication entre l'utilisateur et son ordinateur. Cette communication s'effectue en mode texte : l'utilisateur tape une « ligne de commande » – c'est-à-dire du texte dans le *terminal* – pour demander son ordinateur d'effectuer une opération précise, par exemple rouler un fichier de code R.

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

## 4.5 Pourquoi RStudio ?

### 4.5.1 Qu'est-ce que RStudio ?

Comme plusieurs autres langages de programmation, R est développé grâce à des fonctions écrites par ses usagers. Un IDE, comme RStudio, est conçu pour faciliter ce travail (Verzani, 2011). RStudio est un projet open source destiné à combiner les différentes composantes du langage de programmation R en un seul outil (Allaire, 2011). Il est conçu pour faciliter la courbe d'apprentissage des nouveaux utilisateurs. RStudio fonctionne sur toutes les systèmes d'exploitation, y compris Windows, Mac OS et Linux. En plus de l'application de bureau, RStudio peut être déployé en tant que serveur pour permettre l'accès Web aux sessions R s'exécutant sur des systèmes distants (Allaire, 2011).

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

RStudio facilite l'utilisation du langage de programmation R en offrant de nombreux outils permettant à son 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

#### 4 *R* ou ne pas *R*?

travail, une visionneuse de données et une prise en charge de la coloration syntaxique (Horton, Kleinman, 2015). De plus, RStudio permet de coder dans plusieurs langages et supportent un grande quantité de formats. Il fournit également un support pour plusieurs projets ainsi qu’une interface pour utiliser des systèmes de contrôle des versions tels que GitHub (Horton, Kleinman, 2015).

##### 4.5.2 Avantages et inconvénients de RStudio

RStudio a plusieurs avantages. L’utilisation de l’IDE est facile à apprendre pour les débutants. Les principaux éléments d’un IDE sont intégrés dans une disposition à quatre volets (Verzani, 2011). Cette disposition comprend une console, un éditeur de code source à onglets pour organiser les fichiers d’un projet, un espace pour l’environnement de travail et un quatrième volet où il est possible d’afficher des graphiques ou de la documentation sur différents packages. De plus, on y retrouve la possibilité de créer plusieurs espaces de travail – appelés projets – qui facilitent l’organisation de différents workflows.

Un autre aspect de RStudio que de nombreux programmeurs apprécient est le fait qu’il peut être utilisé via un navigateur Web pour un accès à distance (Verzani, 2011). De plus, l’IDE offre de nombreux outils pratiques et faciles à utiliser pour gérer les packages, l’espace de travail et les fichiers. RStudio supporte plusieurs langages de programmation ainsi que différents langages de balisage. Finalement, de nouvelles fonctionnalités sont souvent ajoutées pour satisfaire aux besoins de la communauté scientifique et le logiciel est régulièrement mis à jour.

Inconvénients : - Peu de configuration, tu peux pas changer les raccourcis, etc - Très limité dans le setup des différents panneaux, tu peux pas voir 2 fichiers en même temps Tu peux pas visualiser 2 fichiers 1 à côté de l’autre, une feature de base dans n’importe quel ide Tu peux pas configurer tes raccourcis clavier, aussi une feature de base Mettons que je veux que ctrl

## 4.5 Pourquoi RStudio ?

+ D copie la ligne, comme dans visual studio, je peux pas - Plus lent que d'autres alternatives pour certaines opérations

Comme plusieurs autres langages de programmation, R est développé grâce à des fonctions écrites par ses usagers. Un IDE, comme RStudio, est conçu pour faciliter ce travail (Verzani, 2011). RStudio est un projet open source destiné à combiner les différentes composantes du langage de programmation R en un seul outil (Allaire, 2011). Il est conçu pour faciliter la courbe d'apprentissage des nouveaux utilisateurs. RStudio fonctionne sur toutes les systèmes d'exploitation, y compris Windows, Mac OS et Linux. En plus de l'application de bureau, RStudio peut être déployé en tant que serveur pour permettre l'accès Web aux sessions R s'exécutant sur des systèmes distants (Allaire, 2011).

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

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

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

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

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

In the lower right are tabbed panes for interacting with the Files (The File Browser), Plots (Graphics in RStudio), Packages (Package Maintenance), and Help system components (The Help Page Viewer). If the facilities

#### 4 *R ou ne pas R?*

are present, an additional tab for version control (Version Control with RStudio) is presented.

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

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

## 4.6 Comment utiliser RStudio ?

La première étape pour commencer à utiliser RStudio est de l'installer<sup>1</sup>. Une fois que cela est fait, ouvrez RStudio. La fenêtre qui apparaît devrait ressembler à l'image ci-dessus. La couleur de l'arrière-plan et celle de la police, la taille des cadrans ainsi que de nombreux autres éléments peuvent être changés. La dernière section de ce chapitre aidera le lecteur à personnaliser son IDE.

### Image de RStudio sans rien, settings de base.

Bien que de nombreux éléments puissent être personnalisés, la disposition par défaut de RStudio est composée de quatre volets principaux (Verzani, 2011). Dans le coin supérieur gauche se trouve le quadran principal. C'est dans celui-ci que l'utilisateur passera la plus grande partie de son temps. On y modifie des fichiers de différents formats et il est possible d'y afficher

---

<sup>1</sup>À cet effet, voir le chapitre 9 du présent ouvrage.



## 4.6 Comment utiliser RStudio ?

des bases de données. Dans le coin inférieur gauche se trouve la console et le terminal. Dans cette première, on peut interagir avec R de la même manière que dans le cadran principal, mais le code ne sera pas enregistré. Le terminal, pour sa part, est le point d'accès de communication entre un usager et son ordinateur. Bien que les différents systèmes d'exploitation viennent avec un terminal déjà intégré, il est aussi possible d'y accéder à partir de RStudio.

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

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

Le cadran dans le coin inférieur droit, pour sa part, contient plusieurs outils très utiles pour les usagers de RStudio. L'onglet *Files* permet à l'utilisateur de naviguer dans les fichiers que contient son ordinateur sans avoir à sortir de RStudio. L'onglet *Plots* permet de visualiser les graphiques générés à partir de R, que ce soit en utilisant *ggplot2*, *lattice* ou *base R*. L'onglet *Packages* permet de consulter les packages installés précédemment par l'utilisateur en plus de pouvoir en consulter la documentation. C'est aussi un des différents endroits à partir d'où il est possible d'installer des packages avec RStudio. L'onglet *Help* permet à l'utilisateur de chercher et de consulter de la documentation sur de nombreux sujets, notamment sur les différentes fonctions en R ainsi que sur les packages.

#### 4 *R* ou ne pas *R*?

Pour sa part, l'onglet *Viewer* permet la visualisation de contenu web local.

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

### 4.7 Personnaliser son RStudio

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

## 5 Baliser les sciences sociales : langages et pratiques

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

Pour ces raisons et plusieurs autres, les chercheurs en sciences sociales font souvent appel aux langages de balisage, ou *markup languages*. Ceux-ci permettent de produire des documents et pages Web sans les limitations des logiciels de traitement de texte. Le présent livre, par exemple, est écrit à l'aide du langage de balisage Markdown et de la plateforme de publication Quarto. D'entrée de jeu, vous vous demandez peut-être quelle est l'utilité d'apprendre ces langages alors que les logiciels de traitement de texte sont nombreux, simples d'approche et en amélioration constante. Ce chapitre tentera donc de répondre aux trois questions suivantes : *Pourquoi apprendre à utiliser des langages de balisage ? Dans quels contextes sont-ils plus utiles que les logiciels de traitement de texte ? Comment les utiliser ?*

L’accent sera mis sur la plateforme Quarto ainsi que sur les langages Markdown,  $\text{\LaTeX}$ ,  $\text{\BibTeX}$  et HTML.

Le premier langage de balisage, le Generalized Markup Language (GML), a été inventé en 1969 par les chercheurs Charles F. Goldfarb, Ed Mosher et Ray Lorie pour la compagnie IBM. Goldfarb et ses collègues devaient intégrer trois applications créées avec des langages différents et avec une logique différente pour les besoins d’un bureau de droit. Même après avoir créé un programme qui permettait aux trois applications d’interagir, ces langages demeuraient différents et avaient chacun leur propre fonctionnement. Le développement de GML a permis de résoudre ce problème en standardisant et en structurant le langage : les mêmes commandes étaient utilisées pour accomplir les mêmes tâches dans chaque programme (Goldfarb 1996). GML a été amélioré durant les décennies suivantes et a été suivi par d’autres langages de balisage, dont  $\text{\LaTeX}$  (1984),  $\text{\BibTeX}$  (1985), HTML (1993), XML (1998), Markdown (2004) et R Markdown (2014).

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

Plusieurs langages de balisage existent et permettent d'effectuer différentes tâches. Le plus répandu est le langage HTML, qui permet de formater des sites web. Le langage XML, lui aussi très utilisé, permet de structurer de larges volumes de données.  $\text{\LaTeX}$  permet pour sa part de formater du texte et de créer des documents en format PDF. Markdown permet également de créer des documents de format PDF, mais aussi HTML et DOCX, contrairement à  $\text{\LaTeX}$ . Le *package*  $\text{\R}$  Markdown permet d'ajouter des extraits de code  $\text{\R}$  à un fichier en langage Markdown. Depuis 2022, le système de publication Quarto permet d'intégrer des extraits de code  $\text{\R}$ , Python,  $\text{\LaTeX}$  ou autre, créés dans différents types d'environnements, à un fichier en langage Markdown.  $\text{\LaTeX}$ , Quarto et Markdown permettent aussi d'intégrer les références bibliographiques du système de traitement de références  $\text{\BIBTeX}$ . Le Chapter 6 explique la manière de citer les références en langage  $\text{\BIBTeX}$  par le biais de Zotero et Better  $\text{\BIBTeX}$ .

Les balises constituent une manière de donner manuellement des commandes au logiciel que vous utilisez. Par exemple, si vous utilisez Microsoft Word, vous avez accès à une panoplie de boutons qui vous permettent de formater votre texte. Les balises exercent les mêmes fonctions, mais de manière manuelle. Lorsque vous appuyez sur un bouton dans Word, celui-ci ajoute des balises au travers de votre texte, mais rend celles-ci invisibles dans l'interface que vous utilisez. Cela permet d'avoir un texte élégant et facile à lire, mais comporte aussi plusieurs inconvénients. Le principal inconvénient est que vous êtes condamné à avoir un pouvoir limité sur le formatage de votre texte. En effet, si les boutons à votre disposition ne vous permettent pas de réaliser une opération, celle-ci sera éternellement impossible à réaliser pour vous. A contrario, les langages de balisage permettent un contrôle presque infini sur les opérations que vous souhaitez réaliser. Incidemment, dans la mesure où vous utilisez le langage approprié pour la tâche que vous souhaitez accomplir, vous devriez être capable de donner exactement la commande nécessaire à votre logiciel. Les langages de balisage, bien qu'ils aient un coût d'apprentissage qui peut s'avérer important et qu'ils soient moins élégants qu'un simple document Word, vous offrent une plus grande flexibilité.

Afin d'utiliser un langage de balisage, il est impératif que le logiciel que vous utilisez puisse prendre en compte ce langage. Un logiciel permet rarement d'utiliser n'importe quel langage. Il est aussi impératif de bien utiliser le langage de balisage. En effet, comme pour les langages de programmation, les langages de balisage ne peuvent pas déduire ce que vous souhaitez leur faire comprendre. Si vous souhaitez mettre du texte en gras, vous devez utiliser les bonnes balises. La moindre erreur est fatale, puisqu'une erreur dans la balise que vous utilisez produira un message d'erreur, le logiciel ne réussissant pas à associer votre balise mal inscrite à une action informatisée. Conséquemment, il est impératif de bien vérifier les balises utilisées afin d'éviter toute erreur qui empêcherait votre document d'être compilé, c'est-à-dire d'être traduit dans son format final<sup>1</sup>. Chaque caractère dans une balise est important et il y a rarement plus d'une seule manière de commander une action. Le positionnement des balises est lui aussi critique : il délimite la portion de texte à laquelle doit être appliquée l'action commandée par la balise.

Il est important de distinguer les langages de balisage des langages de programmation. En effet, ceux-ci sont similaires à certains égards, mais ont des vocations différentes. Les deux s'appuient sur un langage informatisé, mais les langages et leurs objectifs diffèrent. Un langage de programmation définit des processus informatisés alors qu'un langage de balisage permet d'encoder du contenu de manière à ce que celui-ci soit lisible tant pour l'humain que pour son ordinateur.

Dans le contexte de la recherche en sciences sociales, la programmation est généralement utilisée afin de récolter, d'analyser et de présenter visuellement des données. Une fois cartes, tableaux et graphiques produits, ceux-ci peuvent être enregistrés — par exemple en format PDF ou PNG — et inclus au sein d'un document qui sera formaté en utilisant un langage de balisage. De manière simple, le langage de programmation contribue à

---

<sup>1</sup>Les logiciels permettent plus ou moins efficacement d'identifier les balises problématiques. Certains ne produisent qu'un message d'erreur sans donner d'indication sur la source du problème, alors que d'autres ciblent très spécifiquement la ligne de syntaxe où se situe la balise problématique.

l'analyse alors que le langage de balisage est essentiellement utile afin de présenter les travaux de recherche, que ce soit dans un document écrit ou sur un site web. C'est principalement de cette manière que sont utilisés les langages de programmation et de balisage dans le cadre de la recherche en sciences sociales.

## 5.1 Réflexion théorique

La plupart des langages de balisage permettent de remplir l'une des deux fonctions suivantes, qui sont particulièrement importantes dans le contexte de la recherche en sciences sociales : produire des documents écrits et gérer des pages Web. Dans les deux cas, cependant, certains sites Web et applications permettent également de remplir ces fonctions, mais avec des limites importantes.

Pour l'écriture de documents très simples comme une liste d'épicerie ou des notes rapides pendant une conférence, les logiciels de traitement de texte sont tout à fait convenables : ils sont simples et rapides à utiliser, un formatage professionnel du document n'est pas de mise. Utiliser un langage de balisage pour des tâches de base n'est en effet pas nécessaire. Par contre, plus la complexité d'un document augmente, plus il devient difficile d'obtenir un résultat satisfaisant en utilisant un logiciel de traitement de texte tel que Word, Pages ou Writer. A contrario,  $\text{\LaTeX}$  permet de produire des documents de tous les niveaux de complexité, tel que démontré sur la Figure Figure 5.1. Quant à Markdown, sa courbe se situerait logiquement entre celles de  $\text{\LaTeX}$  et de Word. Plus généralement, utiliser un langage de balisage comme  $\text{\LaTeX}$  ou Markdown<sup>2</sup> comporte plusieurs avantages par rapport aux logiciels de traitement de texte traditionnels. Ces avantages peuvent se résumer en quatre concepts : automatisation, personnalisation, flexibilité et qualité graphique.

---

<sup>2</sup>Les avantages et désavantages de Markdown cités dans la prochaine section s'appliquent eux aussi à Quarto. Les avantages ou inconvénients ne s'appliquant qu'à Quarto et non à Markdown tout court sont présentés comme tels.

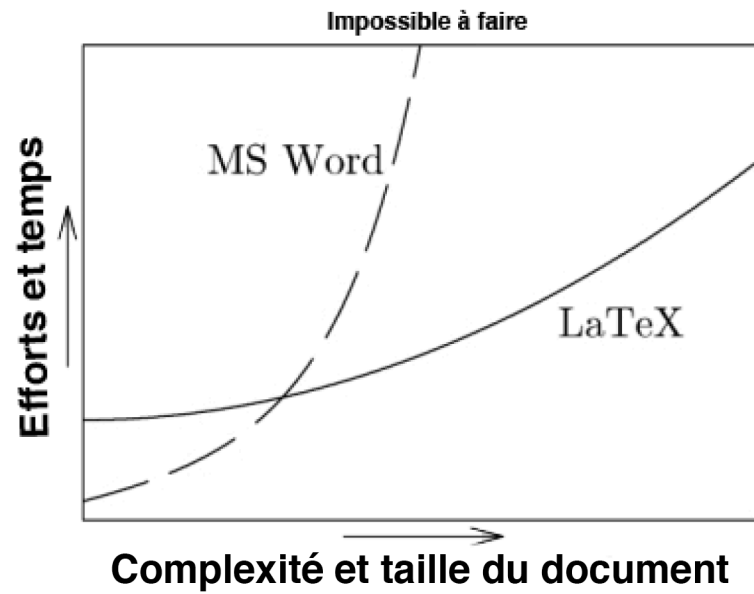


Figure 5.1: Utilité relative de Word et  $\text{\LaTeX}$  selon la complexité et la taille du document

*Source* : Yannick Dufresne (2015).



## 5.1 Réflexion théorique

Premièrement,  $\text{\LaTeX}$  et Markdown permettent d'intégrer une bibliographie *automatique* et professionnelle en utilisant  $\text{\BibTeX}$ . Cette bibliographie peut être adaptée très facilement en différents styles bibliographiques reconnus ou en un style bibliographique personnalisé. Avec  $\text{\BibTeX}$ , plus besoin de vérifier si le titre de l'article est toujours en italique, si le numéro de volume est toujours entre parenthèses ou si le nom de famille des deuxièmes auteurs est toujours avant ou après le prénom puisque tout ceci est fait de manière *automatique*.  $\text{\BibTeX}$  comprend également les différences entre les types de sources (articles scientifiques, livres, sites Internet, etc.) et ajuste leur présentation en conséquence. De plus, si une des sources que vous citez n'est pas incluse dans la bibliographie, une erreur s'affiche, vous permettant d'identifier le problème plutôt que de vous retrouver avec une référence manquante. À l'inverse, si une source est retirée du texte, elle disparaît *automatiquement* de la bibliographie dans le document final mais demeure présente dans le fichier où se trouvent les références bibliographiques. Cela évite les aller-retour pour vérifier que chaque source de la bibliographie se trouve au moins une fois dans le texte et que chaque source dans le texte est citée en bibliographie. Grâce aux balises, en cliquant sur les références incluses dans le document, celui-ci change de page pour se retrouver automatiquement à l'entrée bibliographique associée. Les références  $\text{\BibTeX}$  pour articles scientifiques peuvent être copiées-collées à partir de Google Scholar.  $\text{\BibTeX}$  rend donc extrêmement simple et efficace l'utilisation des références bibliographiques grâce à sa capacité à *personnaliser* et *automatiser* leur présentation.

L'intégration de figures et tableaux dans le texte est aussi rendue très simple et professionnelle grâce à  $\text{\LaTeX}$  et Markdown. La taille de la figure ou du tableau, son positionnement et son intégration par rapport au texte environnant peuvent être réglés de telle sorte que l'ajout de texte avant ou après la figure ou le tableau ne produira pas des résultats inattendus. Au contraire, en définissant des paramètres pour l'ensemble du texte, le chercheur pourra *personnaliser* entièrement la présentation des figures et tableaux. De plus, la qualité des figures et tableaux ne diminue pas lors de leur intégration : les figures restent aussi belles qu'elles l'étaient

originellement, ce qui n'est pas toujours le cas avec certains logiciels de traitement de texte. Les numéros des figures et tableaux sont aussi mis-à-jour *automatiquement*, ce qui veut dire que vous n'aurez jamais à vous préoccuper de modifier leur numéro lorsque vous rajoutez une figure ou un tableau dans le texte. Grâce aux balises, en cliquant sur le numéro associé à la figure ou au tableau dans le texte, le document se retrouve automatiquement à l'endroit où se trouve le graphique ou tableau. De plus, le texte d'une figure PDF peut être surlignable, incluant les titres des axes ou des annotations sur le graphique. Ce type de tâche n'est pas possible à faire à l'aide des principaux logiciels.

Markdown et L<sup>A</sup>T<sub>E</sub>X permettent aussi la gestion *automatisée* de la table des matières, et les références aux pages appropriées se mettent à jour en continu. La table des matières prend en compte l'architecture du texte choisie manuellement par le chercheur, qui est définie par des balises définissant différents niveaux hiérarchiques de sections, sous-sections ou chapitres. Des manières *automatiques* de référencer les figures et les tableaux dans des sections distinctes de la table des matières sont également offertes, encore une fois *personnalisables* au goût du chercheur.

Bien que la mise en page de documents produits via Markdown et L<sup>A</sup>T<sub>E</sub>X puisse être définie entièrement manuellement par un utilisateur expérimenté, les débutants apprécieront les nombreux gabarits (*templates*) disponibles en libre permettent de gérer *automatiquement* la mise en page de manière clé-en-main. Ceux-ci permettent de rendre l'apparence d'un document plus esthétique et uniforme et peuvent être utilisés tels quels ou peuvent servir de point de départ pour un chercheur souhaitant y apporter certaines modifications sans toutefois partir d'une feuille blanche. La majorité des utilisateurs, même les plus expérimentés, utilisent ces gabarits comme base lorsqu'ils rédigent un document. Ceux-ci constituent une mine d'or puisqu'ils rendent accessible le code Markdown et L<sup>A</sup>T<sub>E</sub>X ayant servi à la conception du gabarit, permettant au chercheur de comprendre comment est obtenu le résultat que lui offre le gabarit. Incidemment, le chercheur peut identifier les sections de code produisant certains éléments de mise en page (ex : positionnement des numéros

## 5.1 Réflexion théorique

de page, positionnement du nom des auteurs, etc.) et les modifier ou s'en inspirer afin de modifier d'autres gabarits. L'utilisation de ces gabarits peut s'avérer complexe pour les non-initiés, mais il s'agit d'une complexité qui s'avère ultimement extrêmement productive puisqu'elle permet au chercheur de devenir autonome et d'ajuster les gabarits à sa convenance afin de produire exactement le résultat désiré en terme de mise-en-page. La liste des gabarits disponibles est extrêmement large et ceux-ci peuvent servir une variété de fonctions. En effet, une variété de gabarits professionnels et de haute *qualité graphique* sont offerts gratuitement en ligne pour des articles, des livres, des rapports, des *curriculum vitæ* () ou encore des feuilles de temps pour des contrats rémunérés ().

### *Photos de CVs et feuilles de temps professionnels*

Un autre avantage non-négligeable de Markdown — qui le distingue à cet égard de  $\text{\LaTeX}$  — est la *flexibilité* qu'il offre à ses utilisateurs. En effet, en utilisant Pandoc Markdown, qui est une extension du langage Markdown de base permettant de combiner plusieurs langages de balisage différents en un seul document, il est possible d'intégrer dans un seul document plusieurs langages de balisage différents tels que  $\text{\LaTeX}$ , HTML, CSS ou JavaScript ainsi que du code R en utilisant Quarto. Quarto est également habilité à travailler avec des fichiers Python dans des environnements de type Jupyter Notebook. Ceci permet donc à l'utilisateur de bénéficier des fonctionnalités de différents langages dans un seul document, rendant ainsi possible une variété de *personnalisations* qui ne seraient pas possible autrement. Qui plus est, il est aussi important de noter que Markdown permet de créer des fichiers Word réguliers, PDF professionnels et HTML à partir d'un même document. L'utilisateur peut donc choisir à sa convenance et à tout moment de quelle manière sera compilée le document rédigé. Cette fonctionnalité est particulièrement pratique dans le cadre de collaboration avec des chercheurs n'utilisant pas les langages de balisage ainsi que lors de l'envoi de manuscrits à des revues scientifiques puisque certaines d'entre elles exigent de recevoir ceux-ci sous forme de document Word.

La facilité avec laquelle peuvent être intégrés et gérés les figures et tableaux dans des documents  $\text{\LaTeX}$  et Markdown a déjà été abordée, mais il est important de souligner que l'utilisation de l'extension Quarto permet d'ajouter une couche supplémentaire d'intégration. En effet, Quarto permet de créer une figure grâce à du code R, ainsi que d'intégrer celle-ci au texte et la formater en un seul document. Cela se fait grâce à l'intégration de R *code chunks* dans le document. Le code est produit dans le *chunk* et la figure ou le tableau qui en résulte apparaît dans le document Quarto *et* sur le document fini. La différence entre Quarto et  $\text{\LaTeX}$  est que ce dernier ne peut pas prendre en compte le code R et les figures et tableaux doivent donc être créés dans un document séparé avant d'être intégrés dans le document  $\text{\LaTeX}$ .

Bien que l'apprentissage de  $\text{\LaTeX}$  et de Markdown puisse être parsemé de nombreuses embûches, ces deux langages bénéficient d'une communauté d'utilisateurs en ligne sur laquelle il est possible de s'appuyer afin de résoudre tout problème rencontré. Les utilisateurs — particulièrement les plus expérimentés — sont nombreux à partager leur expérience à leurs collègues rencontrant des problèmes afin de contribuer à régler ceux-ci. Cette communauté est présente sur une multitude de sites Web, bien que le point de rencontre principal soit le forum Stack Overflow (2023), qui est également utilisé pour régler des problèmes de programmation en R. Une simple recherche sur Google d'un problème rencontré avec  $\text{\LaTeX}$  ou Markdown offrira à l'utilisateur des centaines de résultats pertinents afin de l'aider, la plupart de ces résultats étant probablement des échanges sur Stack Overflow. L'utilisateur pourra donc filtrer les résultats et observer les nombreuses solutions envisageables à son problème afin de définir laquelle est la plus appropriée dans sa situation. Il est important de noter, toutefois, que cette communauté est nettement plus développée pour les utilisateurs de  $\text{\LaTeX}$  que de Markdown, puisque ce dernier langage est moins répandu que le premier.

Autre preuve de leur grande *flexibilité* et capacité de *personnalisation*, certaines manières plutôt spécifiques de formater le texte sont présentement uniquement disponibles avec  $\text{\LaTeX}$  ou Markdown. C'est le cas de

## 5.1 Réflexion théorique

la possibilité de séparer automatiquement un mot en deux en fin de ligne à l'aide d'un tiret s'il est suffisamment long, ou encore de la possibilité de permettre que la dernière ligne d'un paragraphe apparaisse seule en haut d'une page — ou, à l'inverse, que la première ligne d'un paragraphe apparaisse seule en bas d'une page. Bien qu'il soit rare que nous ayons absolument besoin de personnaliser le texte de cette manière, ces possibilités peuvent s'avérer utiles lorsque vous rédigez un texte qui doit se conformer en tout point à un gabarit spécifique. En effet, certaines revues scientifiques, maisons d'édition ou universités (dans le cadre de la rédaction de mémoires et thèses) imposent ce type de gabarit inflexible et parfois plutôt capricieux. C'est dans ce type de contexte que la *flexibilité* incomparable de  $\text{\LaTeX}$  et Markdown peut s'avérer utile.

Les langages de balisage permettent également de créer des pages Web. Bien que les pages Web puissent être créées à partir de sites Web comme WordPress, les langages de balisage permettent de produire des résultats plus personnalisables, plus automatisables et avec une plus grande qualité graphique également. Ainsi, l'utilisation de fichiers HTML, JavaScript et CSS, dont le code peut également être intégré dans un fichier Markdown, tout comme c'est le cas avec les fichiers  $\text{\LaTeX}$ , afin de produire des pages Web. HTML peut être appris via des cours sur Code Academy.

Finalement, il est important de mentionner en terminant que Markdown, Quarto et  $\text{\LaTeX}$  sont entièrement gratuits et accessibles aux utilisateurs de tous les systèmes d'exploitation.

Ceci dit,  $\text{\LaTeX}$  comporte quelques difficultés techniques. La plupart d'entre elles peuvent être réglées ou diminuées en travaillant en Markdown et Quarto.

Premièrement,  $\text{\LaTeX}$  est difficile à apprendre. Certaines tâches qui peuvent sembler simples comme l'ajout d'un tableau peuvent nécessiter de nombreuses lignes de code. De plus, à la moindre erreur de frappe dans l'utilisation d'une balise, le code risque de planter et de ne pas produire le document PDF souhaité. C'est ce qu'on appelle une erreur de compilation. La compilation est le processus par lequel un document écrit en langage

de balisage est transformé en fichier textuel, en format PDF dans le cas de  $\text{\LaTeX}$ . Markdown est un langage plus simple à apprendre, avec des balises plus courtes et intuitives. Il occasionne donc moins d'erreurs de compilation.

Deuxièmement,  $\text{\LaTeX}$  est incompatible avec Word, Pages ou Writer. Pour transférer un fichier de traitement de texte vers  $\text{\LaTeX}$ , les balises doivent être ajoutées manuellement une par une. À l'inverse, pour transférer un document  $\text{\LaTeX}$  vers un fichier de traitement de texte, les balises doivent être retirées une par une. Il est aussi possible de copier le texte directement à partir du fichier PDF produit par  $\text{\LaTeX}$ , mais les fins de ligne sont interprétées par Word, Pages ou Writer comme des retours plutôt que des espaces, et les accents sont souvent mal copiés et doivent être réécrits manuellement. Encore une fois, Markdown évite ce problème en permettant d'écrire un fichier DOCX à partir du langage de balisage. Le formatage du fichier DOCX demeure un peu compliqué cependant et doit être fait à partir du modèle d'un autre document DOCX formaté tel que souhaité. De plus, les fichiers DOCX ne peuvent pas être transformés en format Markdown. Quarto permet d'écrire un texte en format Markdown et de produire un fichier DOCX à partir d'un template Word. De plus, pour les fichiers Word à transformer en format Markdown, les balises plus simples en Markdown qu'en  $\text{\LaTeX}$  rendent la tâche plus simple.

Quarto comporte également ses désavantages. Ainsi, bien que plusieurs fonctions  $\text{\LaTeX}$  puissent être utilisées dans des fichiers Markdown, certaines demeurent incompatibles, ce qui rend certaines tâches possibles à faire uniquement par  $\text{\LaTeX}$ . L'inverse n'est pas vrai : les tâches qui peuvent être faites en Markdown peuvent également être faites en  $\text{\LaTeX}$ .

Enfin, il existe des désavantages inhérents à l'utilisation des langages de balisage. C'est pour cette raison que des logiciels comme Word demeurent essentiels pour certaines tâches.

L'un des principaux désavantages de Markdown et de  $\text{\LaTeX}$  est qu'ils ne comportent aucun système de suivi des modifications. Pour réviser un travail fait en langage de balisage, des commentaires peuvent être ajoutés sur

## 5.1 Réflexion théorique

le fichier sortant — nécessairement PDF pour un fichier sortant produit avec  $\text{\LaTeX}$ . Des commentaires peuvent aussi être faits directement dans le document  $\text{\LaTeX}$  ou Markdown, à l’aide de balises spécifiques. Ces commentaires n’apparaissent cependant pas dans le fichier sortant. Le suivi des modifications en  $\text{\LaTeX}$  et Markdown nécessite donc souvent l’utilisation de Git et de GitHub, qui sont abordés plus en détail dans le Chapter 8. Même avec GitHub, les longs paragraphes ayant fait l’objet de plusieurs modifications peuvent être longs à comparer par rapport à Word, qui permet de visualiser les propositions d’ajouts et de retraits de caractères de manière plus intuitive. Le suivi des modifications en Word permet également de distinguer les auteurs de différents commentaires par leurs noms, ce que GitHub ne permet pas de faire. Pour ces raisons, et aussi pour faciliter la mise en page par les éditeurs, certaines revues scientifiques refusent les fichiers PDF et demandent que les soumissions soient faites en format DOCX — ce qui pose problème pour les utilisateurs de  $\text{\LaTeX}$  mais pas ceux de Markdown.

Les langages de balisage comportent également un autre désavantage important : la faible qualité des correcteurs de fautes de français, en particulier les fautes autres que celles d’orthographe. VS Code comprend une extension Grammarly, mais RStudio n’en possède pas une. Le correcteur orthographique est disponible en plusieurs langues, mais ne repère pas les erreurs de syntaxe, de grammaire ou de forme, entre autres. Ces éléments sont pourtant essentiels pour la rédaction de textes académiques. Il est donc souvent utile de faire une révision syntaxique dans Word ou dans Antidote et d’intégrer les corrections en Markdown ou  $\text{\LaTeX}$ .

Somme toute, Word n’est pas à antagoniser et demeure très utile pour des tâches simples. Cependant, dans le monde académique, la production de fichiers de qualité faisant appel à des graphiques, tableaux et blocs de code personnalisés de qualité et automatisés est simplifiée en utilisant des langages de balisage.

## 5.2 Réflexion méthodologique

En pratique, comment utiliser Markdown, L<sup>A</sup>T<sub>E</sub>X et B<sup>I</sup>B<sub>T</sub>E<sub>X</sub>?

L<sup>A</sup>T<sub>E</sub>X a une syntaxe particulière qui demande un certain temps d'adaptation. Pour écrire une phrase simple comme celle-ci, la phrase peut être écrite telle quelle. Par contre, pour mettre un **mot** en caractères gras, il faut utiliser la balise suivante: `\textbf{mot}`. Pour mettre le **mot** en rouge, la balise est `\textcolor{red}{mot}`. Pour le mettre en italique et en note de bas de page<sup>3</sup>, les balises `\footnote{\emph{mot}}` peuvent être utilisées. Ainsi, des balises peuvent contenir d'autres balises. En langage L<sup>A</sup>T<sub>E</sub>X, une balise commence toujours par une barre oblique inversée. Par la suite, le nom de la fonction (*emph*, *textbf*, *textcolor*, etc.) est appelé. Enfin, généralement, le mot à formater est placé entre accolades (`{}`).

Chaque document L<sup>A</sup>T<sub>E</sub>X commence par un préambule. Celui-ci présente des informations telles que la taille des caractères, le type d'article, le format de mise en page, la police de caractères, l'utilisation d'en-têtes et de pieds de page, ainsi que l'utilisation de *packages* L<sup>A</sup>T<sub>E</sub>X permettant différentes fonctionnalités de personnalisation du document.

Il n'est pas nécessaire ni souhaitable d'apprendre l'ensemble des fonctions et des *packages* L<sup>A</sup>T<sub>E</sub>X qui existent. Au contraire, il est souvent mieux de commencer par un gabarit de document qui convient au type de document que vous voulez créer et ensuite de rechercher en anglais sur Stack Overflow la manière d'ajouter des éléments de formatage que vous ne connaissez pas (par exemple, *highlight latex text*). Des gabarits sont disponibles sur Overleaf, au <https://fr.overleaf.com/latex/templates>.

Markdown fonctionne de manière similaire à L<sup>A</sup>T<sub>E</sub>X, mais se démarque par sa plus grande flexibilité et sa syntaxe beaucoup plus légère. Par contre, Markdown est moins *user-friendly*, c'est-à-dire qu'il y est plus difficile de modifier l'aspect visuel d'un document. Tout document Markdown débute

---

<sup>3</sup>*mot*



## 5.2 Réflexion méthodologique

avec un court bloc de syntaxe YAML (acronyme de **Yet Another Markup Language**) qui définit les paramètres généraux du document. Voici un bloc YAML typique :

```
---
title: "Les langages de balisage"
subtitle: "Ça change pas le monde, sauf que..."
author:
  - Alexandre Fortier-Chouinard^[University of Toronto]
  - Maxime Blanchard^[McGill University]
  - Étienne Proulx^[McGill University]
output: pdf_document
documentclass: article
bibliography: references.bib
---
```

Outre le titre, le sous-titre et le nom des auteurs, on y trouve aussi le gabarit servant à construire l'aspect visuel du chapitre, la manière dans laquelle il est compilé — dans ce cas-ci, PDF — ainsi que le chemin d'arborescence afin d'accéder au document  $\text{BIB}\text{T}_{\text{E}}\text{X}$  où sont enregistrées les références utilisées. Il est aussi possible d'y définir la taille de la police ou encore le gabarit servant à définir le type de bibliographie qui sera utilisé. De manière particulièrement importante, c'est l'endroit où sont chargés les *packages*  $\text{L}^{\text{A}}\text{T}_{\text{E}}\text{X}$  qui seront utilisés. En effet, la quasi-totalité des *packages* et fonctions  $\text{L}^{\text{A}}\text{T}_{\text{E}}\text{X}$  sont utilisables dans Markdown, alors que l'inverse n'est pas vrai. Il est donc possible de personnaliser un document Markdown en utilisant des *packages* ayant été créés pour  $\text{L}^{\text{A}}\text{T}_{\text{E}}\text{X}$ .

La syntaxe à utiliser au travers du texte est somme toute plutôt simple. Pour mettre un ou plusieurs **mots en gras**, il suffit de les entourer de deux astérisques (**\*\*mots en gras\*\***); pour les mettre *en italique*, il faut les encadrer d'une seule astérisque (**\*en italique\***). Pour définir un titre de section ou de sous-section, il suffit de mettre des # devant le titre en question. Plus vous ajoutez de #, plus le titre sera petit et plus il sera

considéré à un niveau hiérarchique inférieur dans la structure du texte. La syntaxe Markdown est donc plus légère que celle de  $\text{\LaTeX}$ , dans le but d'en rendre la lecture plus simple pour son utilisateur.

Bien que des gabarits Markdown soient disponibles, ceux-ci sont plus rares. Ils se trouvent pour la plupart sur GitHub, rendus disponibles par leur créateur. Cela étant dit, leur personnalisation peut s'avérer plutôt complexe. En somme, Markdown est particulièrement pratique pour les documents ne nécessitant pas de respecter un gabarit précis et réquérant simplement un document d'allure simple et professionnelle.

Pour sa part,  $\text{\BibTeX}$  a une syntaxe relativement simple. D'emblée, les références  $\text{\BibTeX}$  pour des articles et ouvrages scientifiques sont disponibles sur Google Scholar. Toutefois, pour citer des sites Web ou des articles de médias, la référence doit être écrite à la main selon un format précis. Une bibliographie sur  $\text{\BibTeX}$  peut ressembler à ceci :

```
@book{darwin03,  
  address = {London},  
  author = {Darwin, Charles},  
  publisher = {John Murray},  
  title = {{On the Origin of Species by Means of Natural Selection  
or the Preservation of Favoured Races in the Struggle for Life}},  
  year = {1859}  
}
```

```
@article{goldfarb96,  
  title={The Roots of SGML: A Personal Recollection},  
  author={Goldfarb, Charles F},  
  journal={Technical communication},  
  volume={46},  
  number={1},  
  pages={75},  
  year={1999},
```

```
publisher={Society for Technical Communication}  
}
```

Un fichier `BIBTEX` ne contient rien de plus qu’une série de publications commençant chacune par la balise `@` suivie du type d’article — *article*, *book*, *incollection* pour un chapitre de livre, *inproceedings* pour une présentation dans une conférence, *unpublished* pour un article non publié et *online* pour un site Web sont parmi les plus connus — et des informations sur la publication mises entre accolades. La première information entre accolades est le code de la référence, par exemple `goldfarb96`. Dans le fichier `LATEX`, l’auteur doit écrire `\cite{goldfarb96}` pour voir dans le document PDF compilé Goldfarb (1996); le lien est automatiquement cliquable et renvoie à la notice bibliographique correspondante. L’ordre des publications dans le document `BIBTEX` a peu d’importance, puisque `LATEX` réordonne par défaut la bibliographie en ordre alphabétique.

## 5.3 Trucs et astuces

Où puis-je utiliser ces langages de balisage? Contrairement à Microsoft Word et Apple Pages, plusieurs options

### 5.3.1 Logiciels de bureau

#### 5.3.1.1 MacTeX, MikTeX et autres distributions $\text{\LaTeX}$

#### 5.3.1.2 RStudio (exemples précédents), Visual Studio et autres logiciels du chapitre 5 (tous?)

### 5.3.2 Logiciels en ligne

#### 5.3.2.1 Overleaf

avec les logiciels de bureau qui permettent d'utiliser  $\text{\LaTeX}$  et Markdown, il est impossible de visualiser le résultat final en temps réel. La compilation est nécessaire au préalable.

Il n'y a pas de compteur de mots ou de caractères. Sauf dans Overleaf?

VS Code with LiveShare

Markdown en Overleaf

## 5.4 Références

## 6 La gestion des références

### 6.1 Pourquoi citer ?

Référence à l'idée de la science selon KKV, contribuer à ce qui a déjà été fait.

### 6.2 À quoi sert un logiciel de gestion bibliographique ?

Un outil de référence bibliographique est un logiciel conçu pour aider les chercheurs et chercheuses à gérer et organiser les références bibliographiques de manière efficace. Ces outils sont particulièrement utiles lors de la rédaction d'articles de recherche, de thèses, de mémoires ou d'autres travaux 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 et 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

## 6 *La gestion des références*

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.). Les personnes utilisatrices peuvent insérer des références directement dans leurs documents sans avoir à se soucier des détails de formatage.
4. Collaboration : Certains outils permettent la collaboration en ligne, ce qui permet 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 permet une économie de temps par la centralisation des données sur une 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 qui permet aux utilisateurs d'accéder à leurs références où qu'ils soient. Les sauvegardes régulières assurent que les données ne sont pas perdues en cas de problème technique.
7. Suivi de lecture : Certains outils permettent aux utilisateurs de suivre les articles et les documents qu'ils ont lus, ce qui est particulièrement utile pour garder une trace de la littérature pertinente.
8. Importation et exportation : Les outils de référence bibliographique permettent souvent d'importer et d'exporter des références dans différents formats, ce qui facilite le transfert de données entre différentes plates-formes.

## 6.2 À quoi sert un logiciel de gestion bibliographique ?

En résumé, un outil de référence bibliographique simplifie grandement le processus de gestion des références bibliographiques, de citation correcte et de création de bibliographies cohérentes, ce qui permet aux chercheurs et aux chercheuses de se concentrer davantage sur le contenu de leurs travaux plutôt que sur les détails de formatage.

Avant même de se pencher sur Zotero, bien choisir son logiciel de gestion bibliographique (puisque'il existe de nombreuses possibilités).

Grands avantages (qui sont communs à tous les logiciels de gestion biblio)

**BibTeX** (logiciel auxiliaire de LaTeX) qui permet de : Gérer et traiter des citations et des bases de données bibliographiques personnelles :

Insérer des références provenant d'une ou plusieurs bases bibliographiques ;

Citer des références dans le corps du texte; Créer une bibliographie ou une liste de références dans les documents selon différents styles de citation.

Saisir manuellement ou importer automatiquement des références provenant de bases de données telles que Compendex, Web of Science, Google Scholar, etc. ;

Importer à partir d'autres logiciels de gestion bibliographique comme : EndNote et Zotero.

**EndNote Online** : EndNote est une application gratuite et libre permettant de générer des bases de données personnelles de références bibliographiques. Ces références doivent être saisies manuellement et peuvent provenir de bases de données ou de catalogues de bibliothèques.

Centraliser les références sélectionnées

Citer les références/documents afin de créer des bibliographies inspirées de styles reconnus, tels que APA, MLA, Chicago, etc.

## 6 La gestion des références

Créer un groupe intelligent (Smart Group) afin de regrouper ses sources dans un groupe particulier. Cette fonction permet d'organiser les références de sa bibliothèque et de faciliter son travail de recherche.

Ce qui distingue EndNote des autres logiciels de traitement bibliographique : bien que l'interface soit seulement dispo en anglais, elle s'adapte aux besoins de l'utilisateur ainsi qu'aux exigences de ces travaux. Comme mentionné précédemment, avec ce logiciel, il est possible de choisir le style bibliographique. En revanche, ce qui la distingue réellement des autres logiciels est qu'il est possible de franciser et de modifier manuellement l'interface ainsi que les champs.

**Mendeley** : logiciel de gestion de références et de gestion de documents développé par une société spécialisée dans les informations scientifiques et médicales. En parallèle avec les autres logiciels de gestion de référence présentés précédemment, Mendeley permet d'organiser, de stocker, d'annoter et de partager des articles de recherche ainsi que d'autres types de documents.

- Gestion de références : permet d'importer, d'organiser et de stocker des références bibliographiques provenant de différentes bases de données académiques.
- Permet d'organiser les références en créant des dossiers et des collections en fonction des différents projets de recherche.
- Annoter et surligner des documents PDF, notamment en ajoutant des commentaires, des notes et des surlignages directement sur l'interface de Mendeley.
- Gestion de documents (autres que les références). Outre les références, Mendeley permet de stocker et d'organiser des PDF. Cela facilite l'accès aux documents lors d'une recherche en cours.
- Permet de générer des bibliographies et des citations dans différents styles de format bibliographique.



## 6.2 À quoi sert un logiciel de gestion bibliographique ?

- Les utilisateurs de Mendeley peuvent partager des références, des documents ainsi que des dossiers entre eux. Comme mentionné précédemment, cela facilite la collaboration sur des projets de recherche.
- Enfin, Mendeley offre la possibilité de synchroniser les références, les annotations ainsi que les documents sur plusieurs appareils. Ainsi, cela permet un accès facile et cohérent.

**Carrefour de gestion bibliographique** (site Web du Réseau de l'Université du Québec sur les logiciels de gestion biblio) ? Pertinent ?

Afin de choisir le bon logiciel : page Web qui regroupe toutes les informations concernant l'utilisation des deux logiciels principaux de gestion de référence bibliographique : soit Zotero et EndNote. Il existe également un quiz afin de savoir quel logiciel répond le mieux aux exigences et aux besoins de l'utilisateur.

### Mini conclu

Bien que chaque logiciel à ses spécificités, ils sont tous relativement similaires. C'est-à-dire qu'ils permettent tous d'économiser du temps et de rendre le travail d'équipe plus facile en centralisant les dossiers.

Ainsi, afin de choisir le logiciel qui répond aux besoins ainsi qu'aux exigences de l'utilisateur, celui-ci doit se demander notamment s'il est nécessaire de partager les résultats de ses recherches ainsi que de travailler en collaboration. De ce fait, s'il est nécessaire de partager ses résultats avec le restant de son équipe, l'utilisateur devrait se munir du même logiciel que ses collègues. Enfin, il est important d'utiliser le logiciel que l'utilisateur préfère, comme mentionné précédemment, Zotero et EndNote se ressemblent, mais ils comportent tous des interfaces différentes.

### Transition avec Laurence

**Zotero** (aborder à la fin de cette section du chapitre). Malgré qu'il existe une multitude d'outils, Zotero est le logiciel de gestion bibliographique que nous privilégions au CAPP. Nous expliquerons ces avantages/inconvénients ainsi que son utilisation dans la prochaine section de ce chapitre.

## 6.3 Installation et configuration de Zotero

Dans cette section, vous serez amené à Installer Zotero ainsi que Better Bibtex, un extension de Zotero servant à générer et maintenir à jour des fichiers .bib à partir de Zotero.

### 6.3.1 Zotero

- Installer Zotero
- Installer Zotero Connector
- Une fois Zotero installé, créer un compte Zotero. Prenez note de votre identifiant et partagez le avec Laurence-Olivier pour qu'il vous ajoute sur le groupe Zotero CLESSN.
- Allez dans vos courriels et suivez les directives pour joindre le groupe Zotero CLESSN.

#### 6.3.1.1 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*

### 6.3 Installation et configuration de Zotero

- Une fois l'add-on installé, allez dans les paramètres de Better Bibtex en allant dans l'onglet Zotero > Settings > Onglet Better Bibtex>Open Better Bibtex preferences...
- Dans la section Citation Key Format, collez ceci: `authEtal2.fold.lower.replace(find=".",replace=_+ len + shortyear | veryshorttitle + shortyear`

#### 6.3.1.2 Génération du fichier .bib

Dans Zotero, vous devriez maintenant voir le groupe Zotero CLESSN dans les Group Libraries.

*Il est important de comprendre que tout changement que vous faites dans Zotero sera automatiquement synchronisé avec le groupe Zotero CLESSN. 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.

#### 6.3.1.3 Utilisation de Zotero lors de l'écriture

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

#### 6.3.1.4 Ajouter des références à Zotero

Il y a différentes façon d'ajouter des références au groupe Zotero CLESSN.

## 6 *La gestion des références*

- Drag & drop à partir de votre librairie personnelle
- Drag & drop les pdf que vous avez sur votre ordinateur dans la collection livre-outils. Zotero va essayer de trouver les métadonnées automatiquement.
- Si il ne réussi pas, vous pourrez ajouter la références en cliquant sur la baguette magique en haut à gauche du symbole ” + ” vert. L’outil de baguette magique est utile si vous possédez le DOI ou le ISBN de l’article/livre que vous devez ajouter. Dans les rares cas où Zotero ne trouve rien à propos de votre référence, vous pourrez remplir les différents champs manuellement.
- Utiliser le connecteur à l’intérieur de votre fureteur web. Zotero va aussi tenter de télécharger l’article directement et l’inclure dans la collection approprié.

### **6.3.1.5 Autres raisons d’utiliser Zotero**

- Centralisation des pdf
- Autre

# 7 Une image vaut mille mots

Camille Tremblay-Antoine<sup>1</sup> Nadjim Fréchet<sup>2</sup>

## 7.1 Introduction

Une fois les données collectées, nettoyées, traitées et analysées, une partie centrale du travail d'un scientifique des données est de faire parler les résultats de ses tests empiriques. Il s'agit alors de trouver la meilleure manière de rendre l'information digeste pour les experts et initiés de votre discipline académique ou pour le grand public. La visualisation graphique des données est donc centrale afin de vulgariser les résultats d'une recherche empirique.

L'objectif de ce chapitre est d'apprendre aux codeurs débutants les rudiments de la visualisation graphique en R. Ce chapitre présentera plus particulièrement les packages R *ggplot2* et *dplyr* eux-mêmes téléchargeable à partir du package *tidyverse*. Si *dplyr* permet de préparer les données avant leur visualisation, *ggplot2* est un package dédié à la production de graphiques. Ce chapitre présente sa grammaire avec une série d'exemples (Wickham 2009; Wickham, Çetinkaya-Rundel, and Grolemund 2023).

Ce chapitre est plus technique que théorique et permet aux codeurs débutants d'en apprendre davantage sur la manière de construire des graphiques en R avec des données concrètes. Cependant, la question

---

<sup>1</sup>Université Laval

<sup>2</sup>Université de Montréal

centrale qui devrait vous guider lorsque vous créez des visualisations est la suivante: **Comment optimiser l’intelligibilité des données?** L’objectif d’un graphique n’est pas seulement d’illustrer les données. Un bon graphique devrait permettre de vulgariser une information ou de mettre en saillance un aspect particulier des données. L’objectif communicationnel devrait toujours être gardé en tête. Les graphiques en exemple dans ce chapitre sont construits avec les données de l’Étude Électorale Canadienne de 2019 qui sont facilement téléchargeables sur leur site<sup>3</sup>.

La première section de ce chapitre expose les options et packages également disponibles pour la construction de graphiques en R. La deuxième section de ce chapitre compare les avantages et inconvénients de l’utilisation de *ggplot2* par rapport aux autres packages de visualisation de données qui auront été présentés. La troisième section de chapitre montre des exemples de graphiques construits avec la grammaire de *ggplot2* en utilisant les données de l’Étude électorale canadienne de 2019. Les codes employés pour produire les graphiques en exemple sont disponibles dans l’annexe de ce livre. Ces codes reproductibles permettront aux codeurs débutants d’adapter ces derniers pour leurs propres projets.

## 7.2 Réflexion théorique

### 7.2.1 Les options disponibles

De nombreux *packages* ont été développés dans le langage R dans le but de visualiser des données graphiquement, il devient donc facile de s’y perdre. Heureusement, les options qui s’offrent à nous se précisent lorsque l’on s’intéresse à ce qui est le plus utilisé dans la communauté des codeurs de ce langage de programmation. Les *packages* les plus utilisés représentent des outils qui ont été substantiellement validés et améliorés par leurs

---

<sup>3</sup><http://www.ces-ec.ca/>

développeurs, mais aussi par une importante communauté de codeurs en ligne et de chercheurs universitaires. Trois de ces options sont présentées dans ce chapitre: les graphiques du *Base R*, le *package Lattice* et le *package ggplot2*. Les avantages et inconvénients respectifs de ces trois approches pour la création de graphiques sont explicités dans les sections suivantes.

### 7.2.1.1 Avantages et inconvénients de Base R

Le *Base R* est le langage de base de R et il permet de faire de nombreuses manipulations statistiques sans avoir à installer de *packages* au préalable. Le *Base R* permet notamment de produire des graphiques rapidement. Cela peut être utile pour visualiser la distribution d'une variable ou pour regarder la relation entre deux d'entre elles par exemple. Pour produire un graphique avec le langage de base R, il suffit de faire appel à la fonction `plot()`. Avec la fonction `plot()`, le codeur peut visualiser la distribution d'une variable seule en spécifiant l'axe des  $x$  dans cette dernière. Le codeur peut également visualiser la relation entre deux variables en spécifiant à l'intérieur de la fonction celles qui composeront les axes des  $x$  et des  $y$  du graphique. Les fonctions `barplot()`, `hist()` ou `boxplot()` disponibles dans le *Base R* permettent de spécifier le style de graphique souhaité, qu'on veuille représenter nos données sous forme de diagramme à barre, d'histogramme ou de diagramme en boîtes (Kabacoff 2022, 119–32).

```
# Exemple de graphique avec la fonction barplot() du BaseR

barplot(y, names.arg=x,
  main="Figure 1 - Proportion (%) de répondants par province\n",
  col = "blue",
  sub="\nSource: Étude Électorale Canadienne de 2019
```

Alors qu'un peu tout peut être fait avec le *Base R*, ce langage demeure élémentaire; il est difficile d'innover dans la visualisation ou même de produire des graphiques plus sophistiqués. Le *Base R* peut sembler plus

simple pour l’exploration de données ou pour produire des graphiques de base rapidement, mais ce langage devient rapidement complexe lorsqu’on cherche à améliorer l’esthétique de son graphique ou visualiser des relations entre plusieurs variables, ce que *lattice* et *ggplot2* permettent plus facilement (Wickham 2009, 3–4).

### 7.2.1.2 Avantages et inconvénients de *lattice*

Développé par Deepayan Sarkar, *lattice* cherche à faciliter la visualisation de graphique en facettes. Plus précisément, ce *package* vise à améliorer les graphiques du *Base R* en fournissant de meilleures options de graphisme par défaut pour visualiser des relations multivariées. Ce *package* est donc intéressant pour les chercheurs et les codeurs voulant présenter graphiquement la relation entre plus de deux variables (Kabacoff 2022, 373–77; Sarkar 2008, 2023). Pour produire un graphique de base avec *Lattice*, le *package lattice* doit préalablement être installé dans la bibliothèque de *packages* du codeur et chargé dans sa session au début de son code (voir annexe). Par la suite, le codeur doit spécifier le type de graphique souhaité avec la fonction appropriée<sup>4</sup>. Une fois la fonction choisie, il doit spécifier par une formule les variables x et y ainsi que la troisième variable à contrôler et à visualiser en facettes (*graph\_type(formula / variable en facettes, data=)*).

Si la Figure 1 produite à partir du *Base R* nous permet de visualiser le pourcentage de répondants par province dans l’Étude Électorale Canadienne de 2019, le *package lattice* nous permet de visualiser facilement ce même pourcentage de répondants en tenant compte du positionnement idéologique des Canadiens par province sur l’échelle gauche-droite, comme l’illustre la Figure 2 (0 étant la gauche et 10 la droite).

---

<sup>4</sup>Plusieurs options disponibles comme des histogrammes avec la fonction *histogram()* ou des graphiques de densité avec la fonction *densityplot()*.



## 7.2 Réflexion théorique

```
# Exemple de graphique avec la fonction histogram() du package lattice

histogram(~gaucheDroite | province, data = GraphiqueLattice, breaks = seq(0, 10,
  by = 1),
  main = "Figure 2 - Distribution des Canadiens\n par province sur l'échelle gauche-droit",
  xlab = "\nIdéologie gauche-droite",
  ylab = "Pourcentage (%)\n",
  col = "blue",
  sub="\nSource: Étude Électorale Canadienne de 2019
```

Cependant, le *package lattice* a pour désavantage d’avoir un modèle formel (une grammaire de graphique) moins compréhensible et intuitif que celui de *ggplot2* lorsque vient le temps d’améliorer l’esthétisme des graphiques. De plus, sa plus faible popularité cause que ce *package* reste moins développé par la communauté de codeurs de R que ne l’est *ggplot2*. Nous examinons plus en détail la grammaire de graphique de ce dernier *package* ainsi que ses avantages et inconvénients dans la prochaine section (Kabacoff 2022, 373–77 et 390; Wickham 2009, 6).

### 7.2.1.3 Avantages et inconvénients de ggplot2

Développé principalement par Hadley Wickham, *ggplot2* est un *package R* faisant partie de la collection de *packages* de *tidyverse*. *Ggplot2* peut être donc utilisé avec les autres *packages* centraux de *tidyverse*, limitant ainsi de potentiels conflits entre les fonctions de *packages* qui puissent être incompatibles avec *ggplot2*. Par exemple, le *package dplyr* de *tidyverse* est très utile pour analyser, organiser et préparer vos données à visualiser avec *ggplot2* (Wickham et al. 2019; Wickham, Çetinkaya-Rundel, and Grolemund 2023, 30).

Le principal avantage de *ggplot2* reste sa grammaire qui permet à l’utilisateur de rendre ses graphiques beaucoup plus visuellement attrayants en facilitant la personnalisation esthétique. Ceci permet de

## 7 Une image vaut mille mots

pousser l'esthétisme de vos graphiques à un très haut niveau par rapport aux autres *packages* de visualisation graphique disponibles en R. Les graphiques *ggplot2* se construisent couche par couche, soit par l'ajout des différents éléments du graphique au fur et à mesure dans le code du graphique à construire.

La première couche des graphiques *ggplot* est généralement celle des données et des variables à visualiser. Elle contient plusieurs éléments fondamentaux qui sont essentiels à chaque graphique. Le premier élément est la spécification de l'utilisation du *package ggplot2* qui se fait simplement en appelant la fonction *ggplot2()*. Dans cette fonction, il faut ensuite mentionner quelle est la base de données (*data=*) ainsi que la fonction qui sera utilisée pour positionner les données (*aes()*). Le positionnement le plus courant est de positionner des données *x* par rapport à des données *y*, ce qui se fait de la sorte: *aes(x=, y=)*.

La deuxième couche des graphiques *ggplot2* est celle du *geom*, qui spécifie le type de graphique souhaité. Les types de graphiques les plus couramment utilisés avec *ggplot2* sont les nuages de points (*geom\_point()*), les diagrammes de lignes de tendances ou de séries chronologiques (*geom\_line()*), les courbes de densité (*geom\_density()*) ainsi que les graphiques à bandes (*geom\_bar()*). Mais les possibilités sont infinies (ou presque!) avec *ggplot2* et bien plus de types de graphiques existent.

Les autres couches des graphiques *ggplot2* dépendent souvent du codeur et des étapes de construction de son graphique<sup>5</sup> (Wickham 2009, 77 et 89-93). Le reste de ce chapitre présente la grammaire de *ggplot2* avec un exemple de construction de graphique à bande présenté couche par couche.

```
# Première couche de l'exemple de graphique
# ggplot2 (base de données, variables et geom)
```

---

<sup>5</sup>Les étapes (couches) d'un graphique *ggplot2* ne sont pas nécessairement dans le même ordre d'un graphique à un autre.

## 7.2 Réflexion théorique

```
ggplot(data=GraphiqueExemple, aes(x=province, y=prop)) +  
  geom_bar(stat="identity")
```

Tel que mentionné dans le dernier paragraphe, la première étape est de spécifier la base de données et les variables qu'on souhaite visualiser. Vous vous souviendrez qu'au début de la section nous avons mentionné la collection *tidyverse*, et plus spécifiquement le *package dplyr* qui y est compris. Ce dernier a été utilisé pour nettoyer/calculer la proportion de répondants par province au préalable, ce qui nous permet de positionner directement la variable *prop* dans l'axe y.

7 *Une image vaut mille mots*

### **7.3 Références**

## 8 Outils de recherche: la quête infinie de l'optimisation

Il est tout simplement impossible de se lancer dans l'accomplissement efficace des méthodes proposées dans cet ouvrage sans d'abord se définir une méthode de travail efficace. Peu importe que l'on travaille seul ou en équipe, des dossiers bien classés, une arborescence claire et un stockage sécuritaire sont gages de succès. Un environnement ordonné pour un travail ordonné.

En science sociale,

Question: Où et quand s'arrêter?

### 8.1 Introduction du chapitre

- Il existe une multitude d'outils pour aider le chercheur à optimiser son temps et son efficacité.
- Ce chapitre présentera 2 types d'outils, et proposera une utilisation efficace de ceux-ci:
  1. Logiciel de gestion de communication (Slack et autres) — Jérémie
  2. Logiciel de gestion de projet (Notion et autres) — Maria
  3. Outils de stockage de données (Dropbox, Google Drive, etc.) — Adrien
  4. Logiciel de gestion de versions décentralisé (Git et Github) — Adrien

## 8.2 Logiciel de gestion de communication (Slack)

- Revue non exhaustive des outils disponibles et aperçu historique;
  - Slack
  - Teams (Microsoft)
- Pourquoi utiliser une de ces plateformes
  - La collaboration en recherche est inévitable
  - Même si le ou la professeur.e avec qui vous travaillez ne l'utilise pas, il est certain que plusieurs chercheurs de votre département l'utilise.
  - Les courriels sont dépassés et inefficaces dans un contexte de travail collaboratif.
  - La présence de teamspaces et de chaînes permet de structurer les conversations d'équipe à l'image de vos projets.
  - On garde et retrace plus facilement les conversations relatives à des projets précis.
  - La plateforme permet de regrouper au même endroit les conversations écrites et vidéos.
- Pourquoi choisir Slack
  - Avantages
  - Inconvénients
  - Comment l'utiliser efficacement (en parallèle à Notion?)
    - \* Penser à la structuration des chaînes.
    - \* Faire régulièrement le ménage des chaînes (archivage au besoin).
    - \* Privilégier les conversations dans les chaînes plutôt que dans les conversations privées (à l'exception de certains cas)
    - \* Utiliser les sections pour tirer ses chaînes.
    - \* Épingler des messages et des documents.
    - \*

## 8.3 Logiciel de gestion de projet

- Revue non exhaustive des outils disponibles et aperçu historique;
- Pourquoi choisir Git et Github?
  - Avantages
  - Inconvénients
  - Comment l'utiliser efficacement (en parallèle à Dropbox, etc.)

## 8.4 Stockage de données (Hub, AWS, Dropbox == défi éthique)

- Revue non exhaustive des outils disponibles et aperçu historique;
- Pourquoi choisir Dropbox?
  - Avantages
  - Inconvénients
  - Comment l'utiliser efficacement (arborescence, \_\_sharedFolders, etc.)
  - Mais si on veut avoir un suivi de modifications?

## 8.5 Logiciel de gestion de versions décentralisé

- Revue non exhaustive des outils disponibles et aperçu historique;
- Pourquoi choisir Git et Github?
  - Avantages
  - Inconvénients
  - Comment l'utiliser efficacement (en parallèle à Dropbox, etc.)

## 8.6 Conclusion: un exemple de l'utilisation des outils

- Faire des liens avec l'ensemble du livre!



## **9 Outils d'intelligence artificielle**



## **10 Sapientia Artificiosa: L'Excellence Cognitive au Service des Études Sociales”**



**11 En général, c'est quoi l'IA, le ML, le DL, etc.? (en mentionnant que c'est un champ qui change tout le temps) (JOZEF)**



## 12 La révolution ChatGPT

- Corpus de textes humains donc possibilité de plagiat
- Tracer la ligne entre plagiat et outil d'assistance
- Parler des revues de science po qui veulent qu'on déclare l'utilisation de ce genre de AI
- Brainstorm
- l'API de OpenAI
- Biais de ce genre d'IA





## **13 Deepl, google traduction, etc,**



## **14 Midjourney, génération d'images, etc.**

- Parler du plagiat



**15 1 autre exemple?**



## **16 Machine Learning, reconnaissance d'images etc**

- AWS, Calcul Québec, etc.





## 17 LLM, chatbots, etc.

- Ce qui s'en vient dans ce domaine?

Salut



## 18 Serpents et échelles

Marc-Antoine Rancourt, Flavie Lachance, Justine Béchar, William Poirier

### 18.1 Introduction

Alors que les chapitres précédents se sont consacrés à la présentation théorique et pratique des sciences sociales numériques, le présent chapitre s'efforcera à aider le lecteur à faire sens de la grande quantité d'information que contient l'ouvrage et à commencer sa propre aventure numérique. Ce chapitre peut être vu comme le résultat de la rencontre entre un jeu éducatif et un guide d'apprentissage. À ce propos, le titre n'est pas anodin. Il est possible de visualiser l'apprentissage de nouveaux outils numériques à l'aide d'un jeu de serpents et échelles. Généralement, la progression du protagoniste se fait en toute tranquillité. À certains moments, l'aventure est corsée par des pièges ou des difficultés propres à sa progression – les serpents. Lorsque cela arrive, le protagoniste régresse ou cesse d'avancer. À d'autres moments, l'aventure est facilitée par des occasions positives particulières – les échelles.

En plus d'une représentation visuelle du tableau d'apprentissage de serpents et échelles qui permettra au lecteur de rendre l'acquisition de nouvelles connaissances plus agréable, ce chapitre offre des lectures, des exercices et des travaux pratiques pour les aventuriers numériques débutants, intermédiaires et avancés. Le corps du texte du présent chapitre est divisé en trois parties en lien avec les niveaux de connaissances des outils

numériques – débutant, intermédiaire et avancé – et chacune d’entre elles comportent une première section concernant sur des lectures, une seconde portant sur des tutoriels à faire en ligne et une troisième sur des travaux pratiques à réaliser dans RStudio. De plus, chaque partie termine par un résumé des principaux « serpents » associés au niveau d’apprentissage qui lui est propre. Le chapitre se termine par une section qui aidera le lecteur à voler de ses propres ailes.

## 18.2 Débutant

### 18.2.1 Environnements de programmation

### 18.2.2 Les alternatives à Word : les langages de balisage

### 18.2.3 Serpents

Beaucoup de nouvelles informations ont été présentées jusqu’à présent dans ce livre. Il est normal de se sentir dépasser et de ne pas tout comprendre. En fait, il aurait été surprenant qu’un lecteur qui débute l’aventure numérique ait tout compris. L’important est de garder une attitude propice à l’apprentissage et se rappeler que rien de ceci n’est inatteignable. C’est au tout début du parcours que se trouve le premier serpent : **croire qu’il sera trop difficile d’apprendre, que c’est un objectif impossible à atteindre**. Même les auteurs de ce livre ont, un jour, commencés par faire *Hello World!* dans la console de RStudio. Le premier serpent est souvent lié à un autre piège qui frappe les codeurs débutants : **la peur de demander de l’aide**. Il faut garder à l’esprit qu’une grande quantité des utilisateurs des outils présentés dans le présent livre sont passés par l’incertitude du début et la crainte du jugement des autres. N’ayez pas peur de poser vos questions, c’est comme cela qu’on apprend.

Une autre catégorie de serpents pour les débutants concerne la pratique des connaissances nouvellement acquises. Les serpents de cette catégorie

sont au nombre de trois. Tout d’abord, on retrouve **la croyance qu’il est possible d’apprendre sans pratiquer**. Bien que cela puisse être possible pour quelques personnes ayant une mémoire phénoménale, la réalité est qu’il sera difficile pour le lecteur moyen de retenir l’information contenue dans ce livre et dans les exercices sans pratiquer les nouvelles notions. Le second serpent de cette catégorie est lié à ce dernier point : DataCamp – où il y a des indices et du code déjà écrit – ne forme pas à lui seul des codeurs. Il faut faire attention à **ne pas rester pris dans une boucle infinie de tutoriels**. Faire des tests avec des projets personnels aide à assimiler les nouvelles connaissances en plus d’être plus intéressant.

Le dernier serpent pour débutants est le suivant : **ne pas construire des bases solides avant d’aller plus loin**. Plusieurs nouveaux codeurs, excités par les nouveaux outils qu’ils apprennent, oublient qu’il est primordial de bien comprendre les éléments de base de la programmation et de la gestion de données avant de se lancer dans des projets plus complexes. Bien qu’il ne soit pas requis de connaître la mécanique pour conduire une automobile, il est tout de même parfois utile – voir nécessaire – de comprendre comment entretenir celle-ci.

## 18.3 Intermédiaire

### 18.3.1 La gestion des références

### 18.3.2 Visualisation graphique en R

### 18.3.3 Serpents :

À la suite des différents exercices et lectures complétés dans le cadre de cette familiarisation aux sciences sociales numériques, le lecteur doit s’assurer d’éviter certains pièges qui se dressent sur le chemin des chercheurs de niveau intermédiaire. Le premier d’entre eux est **vouloir apprendre plusieurs langages et n’en maîtriser aucun**.

Plusieurs chercheurs, lorsqu'ils commencent à maîtriser de nouveaux outils, s'emballent et souhaitent en apprendre davantage. C'est une bonne chose, mais il faut faire attention à ne pas apprendre que quelques éléments de plusieurs langages de programmation, et plutôt en maîtriser un. Comme le dit un diction populaire, « qui trop embrasse mal étroit ».

Un second serpent auquel de jeunes chercheurs sont la proie est **coder en n'utilisant pas un style et une planification cohérente et constante**. En n'adoptant pas un style standard – ou en n'utilisant pas le plus souvent le même style – il peut devenir difficile pour les autres et pour soi-même de se retrouver dans le code. Cela peut causer d'importants problèmes de compréhension ou des problèmes techniques. Il est rare qu'un même code ne serve qu'une seule fois. Il est donc de viser à ce que le code qu'on produit soit compréhensible, transférable et – idéalement – optimisé. Un autre serpent s'inscrivant dans la lignée du précédent est **écrire du code mais ne pas le commenter**. Commenter son code contribue grandement à la transférabilité et la pérennité de son travail. Bien que la fonction d'une section de code peut sembler évidente pour son créateur le jour où elle est produite, elle ne le sera pas nécessairement pour d'autres ou pour lui-même dans le futur.

Le dernier piège se dressant sur le chemin d'un chercheur de niveau intermédiaire est de **croire qu'il a suffisamment de connaissances et ne pas sortir de sa zone de confort**. L'apprentissage de techniques plus complexes demande de sortir de sa zone de confort et de se confronter à l'inconnu. Cela demande également d'accepter qu'on ne connaît pas tout et qu'il y aura des échecs et des frustrations. C'est ainsi qu'un chercheur intermédiaire peut dépasser ses limites et devenir un chercheur de niveau avancé.

## **18.4 Avancé**

### **18.4.1 Gestion de projet et de données**

### **18.4.2 Outils d'intelligence artificielle**

### **18.4.3 Snakes**

- La peur de partager son code
- Laisser le parfait être l'ennemi du bien
- Manquer d'empathie et de compréhension envers les nouveaux utilisateurs
- Douchebagisme

## **18.5 Conclusion**





## References

- Adcock, Robert, and David Collier. 2001. “Measurement Validity: A Shared Standard for Qualitative and Quantitative Research.” *American Political Science Review* 95 (3): 529–46. <https://doi.org/10.1017/S0003055401003100>.
- Goldfarb, Charles F. 1996. “The Roots of SGML – A Personal Recollection.” *sgmlsource.com*. 1996. <http://www.sgmlsource.com/history/roots.htm>.
- Kabacoff, Robert. 2022. *R in Action: Data Analysis and Graphics with R and Tidyverse*. Third edition. Shelter Island, NY: Manning Publications.
- Morandat, Floréal, Brandon Hill, Leo Osvald, and Jan Vitek. 2012. “Evaluating the Design of the R Language.” In *ECOOP 2012 – Object-Oriented Programming: 26th European Conference, Beijing, China, June 11-16, 2012. Proceedings*, edited by James Noble, 7313:104–31. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg. <https://doi.org/10.1007/978-3-642-31057-7>.
- Muenchen, Robert A. 2011. *R for SAS and SPSS Users*. 2nd ed. Statistics and Computing. New York: Springer.
- Sarkar, Deepayan. 2008. *Lattice: Multivariate Data Visualization with R*. New York, NY: Springer New York. <https://doi.org/10.1007/978-0-387-75969-2>.
- . 2023. “Trellis Graphics for R.” <https://cran.r-project.org/web/packages/lattice/lattice.pdf>.
- Stack Overflow. 2023. *Stack Overflow*. <https://stackoverflow.com/>.
- Tippmann, Sylvia. 2015. “Programming Tools: Adventures with R.” *Nature* 517 (7532): 109–10. <https://doi.org/10.1038/517109a>.

## References

- Wickham, Hadley. 2009. *Ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer New York. <https://doi.org/10.1007/978-0-387-98141-3>.
- Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy McGowan, Romain François, Garrett Grolmund, et al. 2019. “Welcome to the Tidyverse.” *Journal of Open Source Software* 4 (43): 1686. <https://doi.org/10.21105/joss.01686>.
- Wickham, Hadley, Mine Çetinkaya-Rundel, and Garrett Grolmund. 2023. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. Second edition. Beijing: O’Reilly.