

Outils de recherche en sciences sociales numérique

CLESSN

2023-08-05

Table of contents

Avant-propos

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

Introduction

1 Trois défis pour une contribution aux sciences sociales numériques

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

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

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

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

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